Comparative Genomics of Eukaryotes

ISBN-10: 90-9021390-2
ISBN-13: 978-90-9021390-3

Comparative Genomics of Eukaryotes

Vergelijkende Analyse van Genomics Data van Eukaryoten

Een wetenschappelijke proeve op het gebied van de
Medische Wetenschappen

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. C.W.P.M. Blom,
volgens besluit van het College van Decanen
in het openbaar te verdedigen
op maandag 8 januari 2007
om 15.30 uur precies

door

Vera van Noort

geboren op 26 augustus 1979 te Groningen

Promotor :

Prof. dr. M.A. Huynen

Manuscriptcommissie:

Prof. dr. H.G. Stunnenberg

Prof. dr. F.C.P. Holstege
(Universiteit Utrecht)

Dr. P. Bork
(European Molecular Biology Laboratory)

Voor mijn engelen

Contents

Chapter 1

Introduction

Introduction

In the past decade information on the blue-prints of several hundreds of organisms has been gathered; the complete genomes of small microbes as well as multicellular organisms like us have been sequenced. Expectations of these projects were high, but as it turned out, the genomes on their own were not sufficient to understand the inner workings of the cell. For well studied model organisms the lists of genes could be put in the context of previous knowledge, but even then we had lots of genes for which the function was unknown. And even if the molecular functions of all the individual encoded proteins were known, we would still only have a parts list and no clue how all these parts should be put together to make a complete organism. Nevertheless, the genome sequences played a crucial role in all genome-wide experiments that followed to obtain knowledge on how the parts lists should be put together and to bridge the gap between genotype and phenotype. To gain more insight into the working of the whole cell, biologists started to collect functional genomics data, like mRNA expression levels of all the genes in a time series. In the first genome-scale mRNA measurements (1998) Spellman and co-workers found that yeast genes that showed similar expression profiles throughout the cell-cycle were often involved in the same processes. A second technique that can be applied at a genomic scale is yeast-2-hybrid, which is an assay of physical protein-protein interactions (Ito et al., 2001; Uetz et al., 2000). Physical protein-protein interactions are important for the cell, for example, in signaling pathways that transfer signals from the outside of the cell to adapt levels of transcription. Physical interactions also exist between proteins that are part of larger complexes, Tandem Affinity Purification followed by mass-spectrometry has been applied to discover the composition of all protein complexes in yeast (Gavin et al., 2006; Krogan et al., 2006).

Besides functional genomics experiments, methods have been developed to predict interactions between proteins in prokaryotes from the genome sequences alone. A very powerful method turned out to be conserved gene neighborhood (Dandekar et al., 1998; Overbeek et al., 1999); as prokaryotic genes are organized in polycistronic transcriptional units (the operons) the occurrence of genes together in operons in multiple organisms implies a functional link between these genes. In this thesis, I will extend this concept to eukaryotes by applying evolutionary conservation of interactions derived from functional genomics data within and between organisms. We use these interactions to predict new functions for unknown genes and to predict regulatory elements that potentially regulate these genes. Furthermore, using intelligent data integration, we are able to specify the type of interaction that is predicted. In this introduction, I will explain the methods that have been used in comparative genomics as well as some of the techniques that have been developed in the field of functional genomics. I will conclude with a short summary of the chapters.

Genomics

In July 1995 the first genome sequence of a free living organism was published (Fleischmann et al., 1995); the genome of *Haemophilus influenzae* contained 1709 genes. The first eukaryotic genome, with about 6000 genes, was that of *Saccharomyces cerevisiae*, which was published in 1996 (Goffeau et al., 1996). More and more genomes became available and a draft of the human genome sequence was finished in 2001 (Lander et al., 2001; Venter et al., 2001). After the sequencing of a genome, the first task is to identify the genes. The most straightforward way is to find sequence conservation on the DNA level, which can point to protein coding regions. Sequence similarity is the operationalization of homology: the relation of genes by common descent. Homology between predicted potential proteins is often used to identify the protein coding genes. The alternative is *de novo* gene prediction, which takes into account intrinsic DNA statistics and specific signals in the DNA sequence. After gene identification, the next step is to determine the function of all the predicted genes. Gene function has several aspects, one aspect is the molecular function, like the specific reaction a protein catalyzes or whether it binds to other proteins. If the encoded protein has high sequence similarity with an experimentally characterized protein, it likely performs the same molecular function and may have the same substrate specificity. If the sequence similarity is lower but the proteins are still homologous, it is expected that they perform a similar molecular function, e.g. the same type of reaction is catalyzed but on a different substrate. Homology has been of significant importance for the determination of gene functions, however for many genes there are no homologs with known functions available and different methods have to be developed to discover their functions.

Orthology

Homology of proteins can point to similar functions. Orthology is a subtype of homology and is the evolutionary notion that two genes descend from the same gene in their last common ancestor (Fitch, 1970). It is assumed that genes that are orthologous to each other carry out the same functions. Therefore, orthology is used to transfer the gene annotation from organisms where a gene function is determined experimentally to organisms where the gene function is not known. Moreover, if we want to compare genes between species, it is necessary to define the corresponding genes (orthologs) between the different species.

The bidirectional best hit method

Homology searches like BLAST (Altschul et al., 1990) and Smith-Waterman (Smith and Waterman, 1981) will identify which genes are homologous to each other. The gene in the other organism that is most similar to the query gene is called the Best Hit. Bidirectional Best Hits are found by homology searches carried out in both directions. The Bidirectional Best Hit method assumes that the genes that are most similar to each other in both directions are orthologs. In principal, this method provides a list of candidate orthology relations.

Phylogeny based orthology

Although the BBH method provides a first estimate of orthologous relations, the evolutionary concept of orthology is best approached by phylogenetic inference. Homologs in several completely sequenced genomes are found and after that multiple alignments are made with a multiple alignment program like ClustalW (Thompson et al., 1994). By inferring a phylogenetic tree with a phylogenetic inference method like neighbor joining, orthologous relations can be assigned to genes that according to the tree are the same gene in the last common ancestor. Although this is computationally intensive, we applied this method to identify orthologs between the yeast *S. cerevisiae* and the worm *C. elegans* in chapter 2.

Inparalogs

One process that hampers orthology detection is gene duplication. Gene duplications are ubiquitous in eukaryotes. Therefore, an extension to the BBH method includes recent duplications. The INPARANOID (O'Brien et al., 2005) program assigns orthologous relations to bidirectional best hits and the paralogs that are more similar to the query gene than to the gene in the other organism. The BBH method with inparalogs can be extended to multiple species by taking triangles of Bidirectional Best Hits including inparalogs and connecting all triangles that share one side (or one BBH). This is implemented in the COG procedure (Tatusov et al., 2001). The COGs are Clusters of Orthologous groups and are manually curated after automated identification. In chapter 4, results of conserved interactions are compared between BBH orthology and COG orthology.

Comparative genome analysis

In order to predict the functions of all the genes encoded in a genome, comparative genome analysis has exploited the genomes themselves to predict functional links between genes. A functional link means that the involved genes are acting in the same biological pathway or process or are part of the same protein complex. Functional links can be used to predict functions for previously unknown genes.

Co-occurrence

The most general indication for a functional link between two genes is the co-presence and co-absence over a large number of genomes (Huynen and Bork, 1998; Pellegrini et al., 1999). Statistical significance can be scored by mutual information of the two phylogenetic profiles (Huynen et al., 2000). Adjustments to the original definition have been made to include phylogenetic information, e.g. clades of species are taken as one species if they all have the same presence-absence pattern of a pair of genes (von Mering et al., 2003). The co-occurrence of the gene encoding frataxin with several genes with known functions was used to predict the involvement of frataxin in iron-sulfur cluster assembly (Huynen et al., 2001). This prediction was later verified by experiments.

Gene-neighborhood

About fifty percent of prokaryotic genes are organized in polycistronic transcriptional units, meaning that several genes are transcribed from the DNA in a row on one unit of mRNA.

The cell can use these operons to transcribe genes that are involved in the same biological pathway or protein complex at the same time. Due to extensive shuffling, genes that are neighbors in one genome are usually not neighbors on another, distantly related genome (Mushegian and Koonin, 1996). If genes occur together in operons in more than one genome, apparently there is a selection pressure to co-transcribe these genes. This gene neighborhood conservation is very useful to predict functional links between genes (Dandekar et al., 1998; Overbeek et al., 1999) and the complete network of these links has been used to identify functional modules (Snel et al., 2002), i.e. groups of genes that act in the same pathway or are part of the same complex. The gene-neighborhood concept can only be used in prokaryotes due to the absence of operons in eukaryotes.

Gene fusion

The third type of genomic context is the occurrence of genes as separate entities, and, in another genome, as a single fused gene (Enright et al., 1999; Marcotte et al., 1999). Gene fusion is the event that two protein coding genes become one gene and code for one polypeptide. This is a strong signal that the gene products are physically interacting in the organisms where they have not been fused. Unfortunately, gene fusions do not occur very often and therefore they are quantitatively not so powerful as a tool to identify functional links (Huynen et al., 2000).

Functional genomics

When the complete genome of an organism is known it can be used to elucidate protein functions by performing functional assays for all genes at the same time. We call this functional genomics (Fields et al., 1999). Whereas in the pre-genome era biologists were often focusing on one gene or one system, with functional genomics it is possible to get an unbiased and comprehensive, albeit one-sided, view of biological processes.

mRNA chips

One aspect of gene function is gene expression, which can be measured by an mRNA chip or microarray. Such a chip is actually a glass slide with specific short single-stranded DNA strings printed on it (Schena et al., 1995). Once a genome has been sequenced, it is possible to make a tiling array, which is a genome chip where every part of the genome sequence is represented by such a short DNA string. A genome chip can be used to measure the transcription levels of all the genes encoded by a genome at once. First, the mRNA is fluorescently labeled. As mRNA is also single stranded it will hybridize to the corresponding DNA string on the chip. After washing away all the unbound mRNA, the fluorescence is measured for each position on the chip. The corresponding position on the genome is known and the mRNA levels of all the genes are now also known. If a specific condition like stress is applied to a cell, the mRNA levels of the genes that are involved in adaptation to stress are expected to change. By measuring the mRNA levels of the genes before and after stress-induction, the stress-responsive genes can be identified (Richmond et al., 1999). Hughes and co-workers have measured levels of mRNA with genome chips after deleting single or multiple genes in yeast (Hughes et al., 2000). In this assay there were no conditions (like stress) to which

genes respond that could be used to predict gene functions. However, it is possible to identify groups of genes that respond in the same way to all deletions. If one gene in the group has a known function and is involved in a specific process, it is expected that other genes of the group are involved in the same biological process. Unfortunately, this method is not very reliable for functional annotation. In chapter 2 we will show that, similar to what was shown earlier for the conservation of co-expression by means of operons in prokaryotes, evolutionary conservation of co-expression in eukaryotes can be used for the prediction of reliable functional links. Large-scale mRNA expression measurements were also done in the malaria causing parasite *Plasmodium falciparum*, specifically in the intra-erythrocyte life stages. In chapter 4 we will use these data to predict new functions for hypothetical *Plasmodium* genes and in chapter 5 we will use the co-expression to predict regulatory elements in upstream regions of *Plasmodium* genes.

ChIP-on-chip

Another way to address gene regulation, that actually also relies on genome chips, is to determine the location of DNA binding proteins on the genome. The ChIP-on-chip procedure starts by cross-linking all proteins that bind to the DNA (Figure 1.1). The DNA is then fragmented by sonication, and an antibody is used to Immuno Precipitate all the copies of the protein of interest. The cross-linked DNA will precipitate with the proteins, hence the name Chromatin Immuno Precipitation (ChIP). The proteins are released from the DNA and the DNA fragments are fluorescently labeled and hybridized to the genome chip (ChIP-on-chip). The fluorescence is measured and locations of protein binding can be mapped to the genome. Lee and co-workers (2002) have analyzed more than a hundred transcription factors in yeast to find their binding locations. They have shown that the genes neighboring the binding sites are often regulated by these transcription factors.

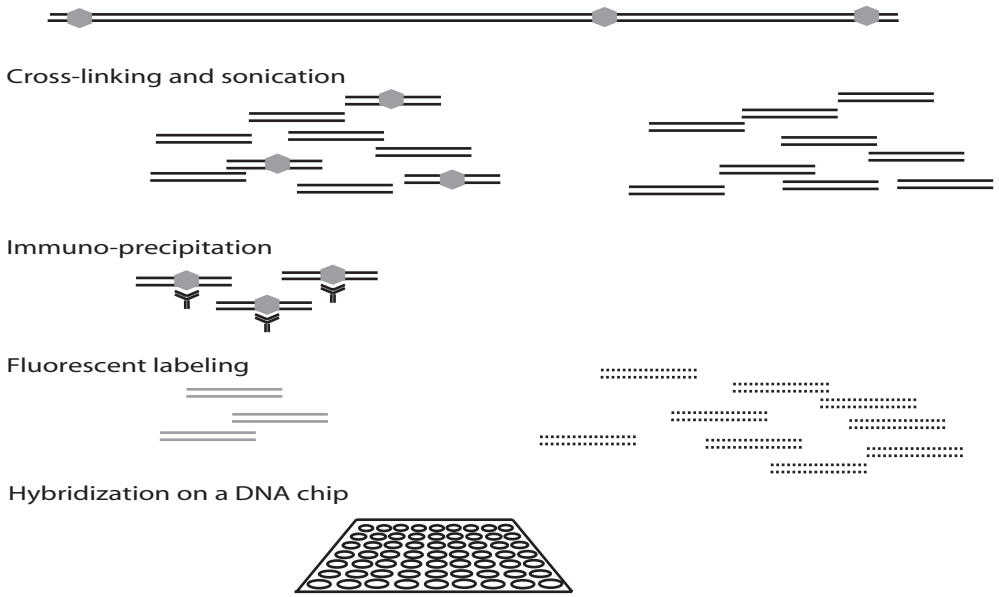


Figure 1.1 | ChIP-on-chip stands for Chromatin Immuno Precipitation followed by a chip experiment. First the proteins are cross-linked to the DNA and the DNA is fragmented into smaller pieces by sonication. With a specific antibody the proteins and the DNA that is bound to them are Immuno Precipitated. The precipitated DNA is fluorescently labeled (left) in a different color than a sample of whole genome DNA (right). Binding locations of proteins are found by identifying spots where the ratio of fluorescence of the immuno-precipitated DNA to the fluorescence of whole genome DNA is larger than one.

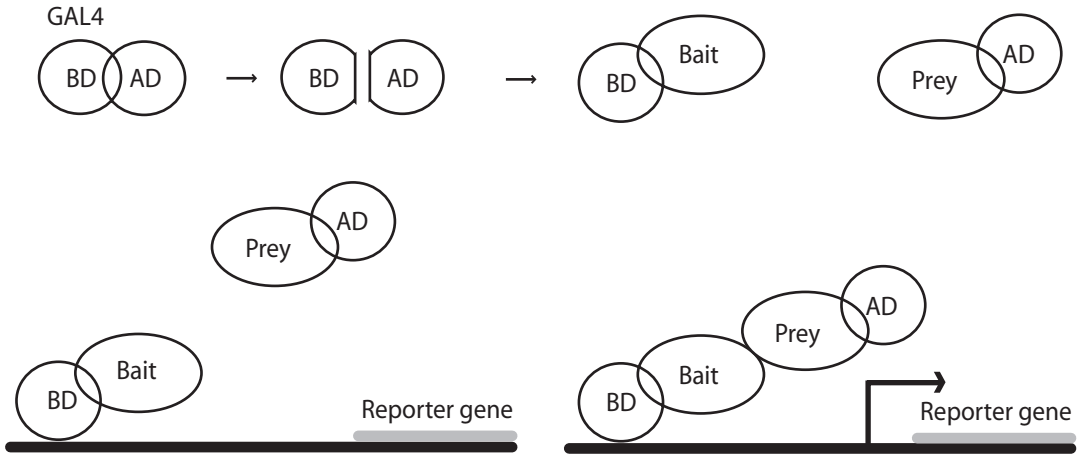


Figure 1.2 | The yeast-2-hybrid system. The transcription factor GAL4 consists of a binding domain (BD) and an activation domain (AD) that can be split. The domains are fused to the bait and prey proteins respectively. If the bait and prey proteins do not bind, the reporter gene stays inactive (left). If the bait and prey protein bind to each other, the reporter gene is transcribed (right).

Synthetic lethality

To predict functional links, synthetic lethality of pairs of genes can also be used (Kelley and Ideker, 2005). Most deletions of genes are not lethal to an organism, but induce a phenotype where the cell is more sensitive to specific conditions. It can also be that two gene deletions are on their own not lethal, but if they are combined the organism dies. This feature is called synthetic lethality. One can imagine that if two genes are part of complementary pathways, deleting both genes will cause the organism to die. A number of large-scale studies have been done, that unfortunately do not encompass all pairs of gene deletions. By identifying genes that share synthetic lethal partner genes, we can find pairs of genes that are part of the same pathway, or actually, as we will show in chapter 7, are part of the same protein complex.

Proteomics

More direct assays of protein function are performed by proteomics techniques. Proteomics is the large-scale study of proteins, particularly their structures and functions (Anderson and Anderson, 1998). This term was coined to make an analogy with genomics, and while it is often viewed as the "next step", proteomics is much more complicated than genomics. Most importantly, while the genome is constant within one organism, the proteome differs from cell to cell and is constantly changing. Most high-throughput proteomic techniques are based on mass spectrometry. The levels of protein expression can be measured by the number of peptides identified that uniquely belong to that protein. We can treat these expression levels in a similar way as mRNA expression data. As we will show in chapter 4, co-expression on the protein level can also be used to predict functional links between proteins.

Yeast-2-hybrid

The interactions between proteins are important for many biological functions. For example, signals from the exterior of a cell are mediated to the inside of that cell by protein-protein interactions of the signaling molecules. Two-hybrid screening is a technique used to discover protein-protein interactions by testing for physical interactions between two proteins (Fields and Song, 1989). One protein is termed the bait and the other is the prey. The concept behind the test is the activation of a reporter gene by the binding of a transcription factor (GAL4) onto the reporter promoter (Figure 1.2). A transcription factor is split into two separate domains; the Binding Domain binds to the promoter and the Activating Domain activates transcription. Even though the transcription factor is split into two fragments, it can still activate transcription when the two fragments are indirectly connected. Bait proteins are fused with the Binding Domain, whereas preys are fused with the Activation Domain. If the bait and prey proteins interact, the two domains of the transcription factor are indirectly connected and transcription of the reporter gene occurs. Two large-scale yeast-2-hybrid screens have been performed covering almost all yeast proteins (Ito et al., 2001; Uetz et al., 2000) and one screen for the fruit fly (Giot et al., 2003). Screens of proteins from other organisms are on their way.

Tandem Affinity purification

Proteins that interact in a yeast-2-hybrid assay may in "real life" never be at the same time at the same place in the cell and may therefore never truly interact. Unlike yeast-2-hybrid, affinity purification methods are well suited for studying complexes under near-physiological conditions. They allow proteins that are fused with a tagged bait to be immuno precipitated with the bait. The proteins that bind to the tagged protein are also retrieved and can be identified by mass spectrometry. These methods have been applied as large-scale screens in prokaryotic and eukaryotic cells. Recently, two groups have published genome-wide screens for protein complexes in yeast (Gavin et al., 2006; Krogan et al., 2006).

This thesis

This thesis presents a number of bioinformatic analyses that cover comparisons between different types of data obtained by functional genomics and proteomics techniques (Figure 1.3) with the aim of both predicting (specific) functional links between genes and studying their evolution as well as finding regulatory elements in non-coding DNA.

Chapter 2 describes the conservation of co-expression after gene duplication and speciation. Based on the notion of conserved operons in prokaryotes, we hypothesize that conservation of co-expression in eukaryotes will point to gene pairs that are functionally linked. First, we show that there is a small but significant amount of co-expression conservation after parallel gene duplication and speciation. The apparent lack of conservation seems to depend both on spurious co-expression and on rapidly evolving, regulatory interactions. Secondly, we show that in case of conservation of co-expression, the gene pairs tend to be involved in a similar biological process. We describe two case stories where a function could

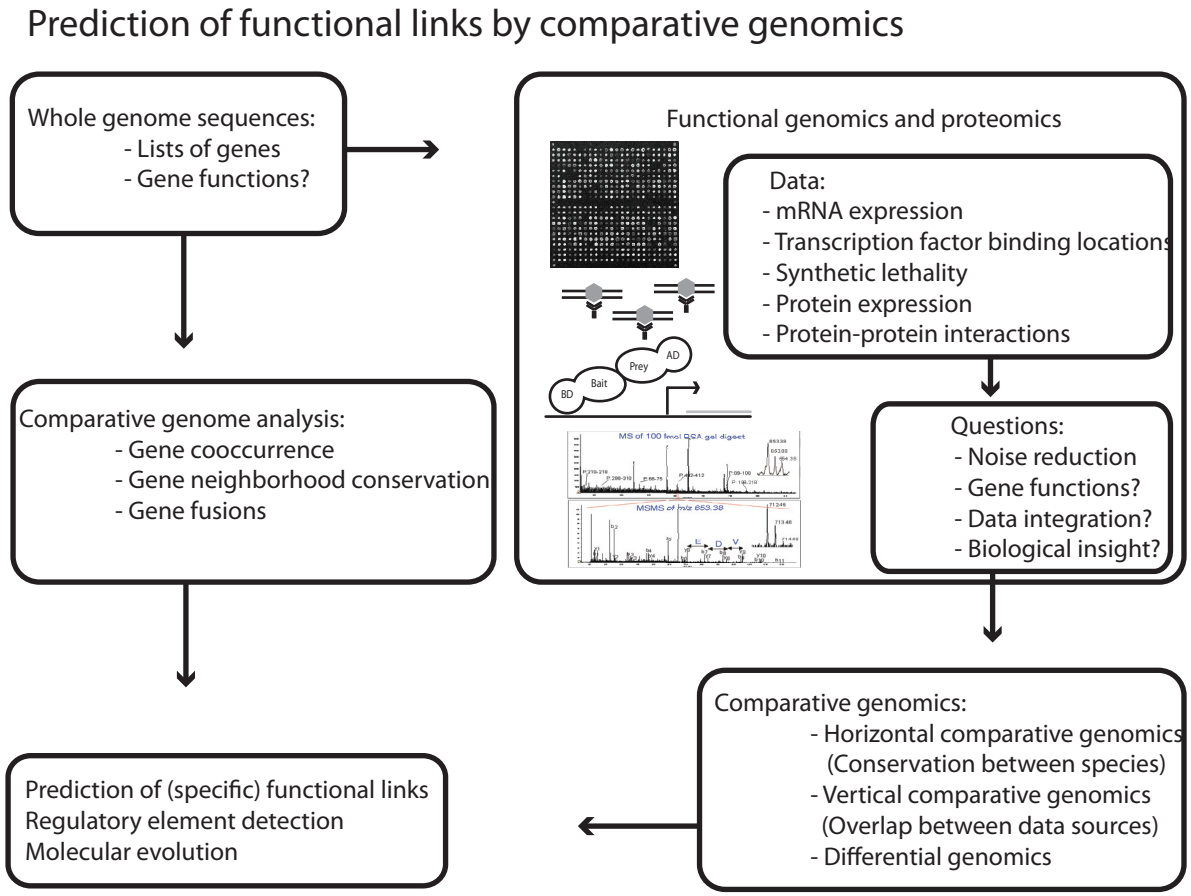


Figure 1.3 | Prediction of functional links by comparative genomics. After the sequencing of a genome, a list of genes is available to which functions have to be assigned. Comparative genomics has proven to be able to predict functional links between genes and has allowed the prediction of functions for hypothetical genes that are linked to genes with previously identified functions. Whole genome sequences have also provided the opportunity for comprehensive biological experiments, i.e. measurements for all genes at once. Functional genomics has provided us with huge amounts of data that in principle can be used to find functions for unknown genes. However, the data is notoriously noisy and the gain of biological insight is not straightforward. In this thesis horizontal comparative genomics (conservation between species) and vertical comparative genomics (overlap between different types of data) methods are developed to reduce noise in functional links derived from functional genomics and proteomics data. A method is developed for regulatory element detection based on groups of genes with reliable functional links. Furthermore, we use differential genomics to predict the type of functional link that exists between genes.

be predicted to a previously unknown gene, using the conserved co-expression.

Chapter 3 presents the network of gene co-expression in yeast. Genes are presented as nodes of a network connected if they are co-expressed. The network statistics show that this is a scale-free, small-world network. The small-world phenomenon means that if two nodes are connected two a third node, the likelihood that they are connected to each other is high. Furthermore, despite the high level of clustering the number of nodes one needs to cross to get from any node to any other node in the network is very small. Scale-freeness means that the number of connections per node follows a power-law distribution. By simulating an evolving genome with gene duplications and mutations and duplications of regulatory elements we show that such a network topology can be obtained without the influence of selection on its large-scale properties. This implies that the global network topology can be explained without invoking selection.

In chapter 4 we describe several methods to combine functional genomics data to obtain reliable predictions of gene functions. If a yeast-2-hybrid interaction is not only observed in one screen but in two independent screens it turns out to be much more reliable. The same holds if the screens were performed in two distantly related species, yeast and fly. It even turns out that the overlap of interactions between two screens within yeast is not much higher than the overlap of interactions between two species. This leads us to propose that physical interactions are highly conserved in evolution. We also show that if not only one but three different functional genomics data sets suggest an interaction between two proteins in *P. falciparum*, the likelihood that this truly is an interaction is much higher than if the interaction is predicted by only one of the genomics data sets.

In chapter 5, we cluster *P. falciparum* genes according to similarity in mRNA expression in two functional genomics screens. We use the upstream regions of these gene clusters together with the upstream regions of their orthologs in *P. yoelii* to find regulatory elements in the genome of *P. falciparum*. We show that some of these elements are indeed correlated with certain expression patterns. Moreover, it appears that most genes of *P. falciparum* contain more regulatory elements in their upstream region than a eukaryote with a comparable number of genes like yeast. In conjunction with the paucity of transcriptional regulators in *P. falciparum* we propose that this organism uses few regulators in a combinatorial fashion, to obtain differential expression patterns..

Predictions of functional interactions from genomics data are usually very general. They do not predict the type of interaction two proteins might have. In chapter 6 we develop a method to distinguish metabolic from physical interactions. First, we show that different functional genomics methods can be indicative of different types of interactions. Smart combinations of data types can lead to extraction of only metabolic interactions.

Chapter 7 provides a summarizing and synthesizing discussion of the chapters presented in this thesis. Some new developments and future perspectives are discussed.

Chapter 2

Predicting gene function by conserved co-expression

Vera van Noort, Berend Snel and Martijn A. Huynen

Trends in Genetics 19, 238-242 (2003)

Predicting gene function by conserved co-expression

Abstract

We show that gene co-expression, which generally provides only a very weak signal for the prediction of functional interactions, can provide a reliable signal by exploiting evolutionary conservation. The encoded proteins of conserved co-expressed gene pairs are highly likely to be part of the same pathway not only after speciation (98%), but also after parallel gene duplication (97%). Conserved co-expression combined with homology data enables us to predict specific gene functions. The use of conservation between parallel duplicated gene pairs to predict function is especially promising given that gene duplication is common in eukaryotes, and that data from only a single organism can be used.

Introduction

One of the major goals of the post-genomic era is the elucidation of gene function. Correlations between expression patterns (Eisen et al., 1998) from hundreds of experiments for both *Saccharomyces cerevisiae* (Hughes et al., 2000) and *Caenorhabditis elegans* (Kim et al., 2001) can predict only general functional interactions (Noordewier and Warren, 2001; Wu et al., 2002). As the evolutionary conservation of weak signals (like gene order), has been used successfully to predict gene function (Huynen et al., 2000; Overbeek et al., 1999), here we examine whether the conservation of co-expression can be used to improve function prediction. We use conservation between pairs of orthologs in two species, as well as conservation of co-expression between parallel duplicated gene pairs in one species to predict functional interactions. We combine these predicted interactions with homology data to predict specific functions for uncharacterized genes.

Co-expression provides a weak signal for pathway prediction

Two large-scale expression datasets were obtained, one from *S. cerevisiae* (Hughes et al., 2000) and one from *C. elegans* (Kim et al., 2001). Uncentered correlation (Eisen et al., 1998) was calculated between the expression profiles of all S. cerevisiae genes and between the expression profiles of all *C. elegans* genes. The higher the correlation (r) between two genes, the more probable it is that they act in the same pathway (Fig. 2.1). However, at a significant correlation threshold of 0.6 ($P<0.005$, Table 2.1) , the fraction of annotated proteins that are part of the same pathway is only 54% in *S. cerevisiae* and 34% in *C. elegans*.

Significant levels of evolutionary conservation of co-expression

To evaluate whether evolutionary conservation (Fig. 2.2) can improve upon these limits in the use of co-expression for function prediction, we first established whether there is sig-

Table 2.1 | Significant levels of co-expression conservation after gene duplication or speciation

	Total pairs[a]	Nr of pairs > 0.6[b]	Observed > 0.6[c]	Expected > 0.6[d]	Observed/expected
Gene pairs with an orthologous gene-pair > 0.6					
C. elegans	18161	803	0.0442*	0.00379	12
S. cerevisiae	36548	1215	0.0332*	0.00216	15
Gene pairs with a paralogous gene-pair > 0.6					
C. elegans	207214	29031	0.1401*	0.00379	37
S. cerevisiae	38253	2167	0.0566*	0.00216	26
Gene pairs with a diverged paralogous gene-pair > 0.6					
C. elegans	125852	1299	0.0103*	0.00379	3
S. cerevisiae	26941	174	0.0065*	0.00216	3

[a] The number of gene pairs, regardless of their co-expression, with a co-expressed, orthologous gene pair in the other species or a co-expressed paralogous gene pair in the same species.

[b] The number of co-expressed gene pairs with a co-expressed, orthologous gene pair in the other species or a co-expressed paralogous gene pair in the same species.

[c] Observed fraction of conserved co-expressed pairs. Asterisk shows $P < 0.001$, determined by 1000 Monte Carlo simulations; that is, such high levels of conservation were not observed when randomly distributing the correlations over the gene pairs 1000 times.

[d] Expected fraction assuming no conservation of co-expression, determined by the total fraction of co-expressed gene pairs from the total number of gene pairs in that species.

nificant conservation, potentially reflecting selection pressure on maintaining functional interactions. To determine conservation between *S. cerevisiae* and *C. elegans*, we first need to define which genes are orthologs of each other, which we do based on phylogenetic trees allowing for multi to multi orthology relations (Fig. 2.3). We found 18161 *C. elegans* gene pairs that have an orthologous pair in *S. cerevisiae* with a co-expression correlation higher than 0.6. Of these, 803 also have a correlation higher than 0.6 in *C. elegans* itself (Table 2.1). Defined this way, 4.4% of the co-expression is conserved, which is 12 times higher than expected assuming no conservation of gene co-expression. Vice versa, of the *S. cerevisiae* gene pairs that have an orthologous pair in *C. elegans* with a correlation higher than 0.6, 1215 also have a correlation higher than 0.6 in *S. cerevisiae* itself, which is 15 times higher than expected (Table 2.1).

Although significant ($P<0.001$, determined by 1000 Monte Carlo simulations), the observed level of conservation of co-expression between S. cerevisiae and C. elegans is quite low (Table 2.1) as already reported (Teichmann and Babu, 2002). However, given that at correlations higher than 0.6 in a single species there are still many false positive predictions, this apparent lack of conservation might be due to spuriously detected co-expressed genes. Consistent with this, genes with a high co-expression correlation in *C. elegans* ($R>0.9$), which we expect to be truly co-regulated, are often co-expressed in *S. cerevisiae* (55%, $R>0.6$). Interestingly, a considerable fraction (50%) of the gene pairs that have co-expression correlation higher than 0.9, but are not conserved ($R<0$ in the other species), encode regulatory proteins. They include a TATA-binding protein (T20B12.2) that is co-expressed in *C. elegans* with a ring-type zinc-finger protein (EEED8.9), and in *S. cerevisiae* an RNA-binding protein (YOR319W) with a protein containing a BAF60b domain (YOR295W) that facilitates the function of transcriptional activators. The lack of conservation appears therefore to depend both on spurious co-expression and on rapidly evolving, regulatory interactions.

Next we determined conservation of co-expression between gene pairs within a species

after parallel gene duplication (Fig. 2.2). The number of such pairs is actually higher than the number of pairs with co-expression conserved between species: 29031 in *C. elegans* and 2167 in *S. cerevisiae* (respectively 37 and 26 times higher than expected; $P<0.001$). Conservation of co-expression within duplicated gene pairs coupled to divergence between the pairs, was studied by selecting the pairs A–B and A'–B' where the correlations between A and B, and between A' and B' are both higher than between A and A', and between B and B'. This conservation is lower than without divergence, but still higher than expected (Table 2.1). Thus, even after differentiation in expression pattern, there is significant conservation of co-expression.

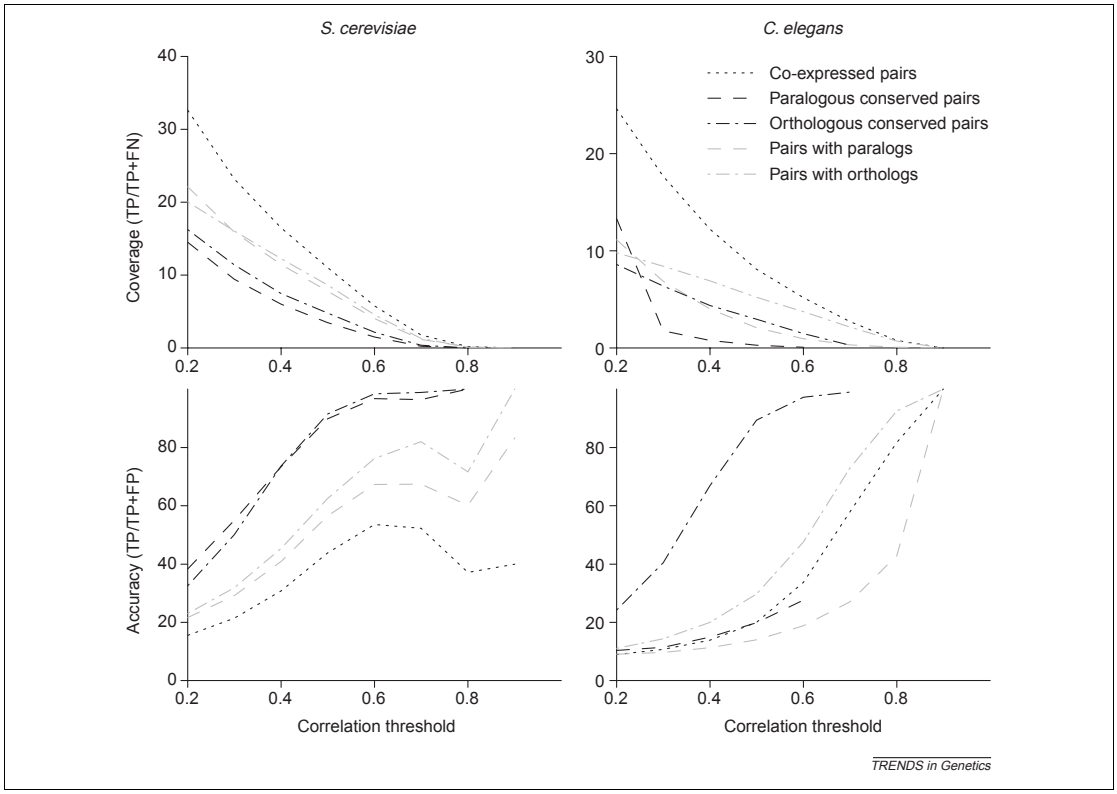


Figure 2.1 | Accuracy and coverage of functional interaction prediction. **Accuracy (bottom) and coverage (top) at varying correlation thresholds for detection of co-expression. The accuracy is obtained by the number of predicted pairs that are on the same map in the PATHWAY database of KEGG (release 23) [23] (true positives) divided by the total number of predicted pairs (true positives plus false positives). The coverage is obtained by dividing the true positives by the total number of gene pairs that can be found on the same map in the PATHWAY database (true positives plus false negatives). Left, co-expressed gene pairs in Saccharomyces cerevisiae. Right, co-expressed gene pairs in Caenorhabditis elegans. Dotted lines, all gene pairs with expression correlation above the threshold; grey dashed lines, gene pairs with expression correlation above the threshold and a pair of paralogs in the same species; grey dot-dashed lines, gene pairs with expression correlation above the threshold and a pair of orthologs in the other species; black dot-dashed lines, co-expressed gene pairs with expression correlation above the threshold and orthologs with an expression correlation above the threshold; black dashed lines, co-expressed gene pairs with expression correlation above the threshold and paralogs with an expression correlation above the threshold. The increased accuracy of conserved co-expression is partly due to the requirement that both genes to have an ortholog in the other species or a paralog in the same species: the accuracies for gene pairs with orthologs or paralogs are slightly higher than the accuracies for all co-expressed gene pairs, although they fall well below the accuracies attained for conserved co-expressed gene pairs.**

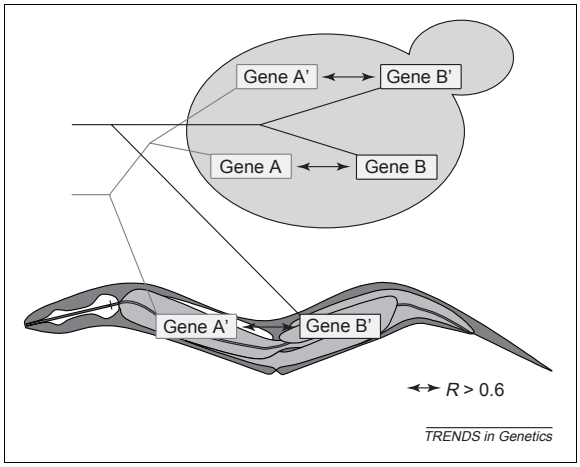


Figure 2.2 | Conservation of co-expression after gene duplication or speciation. Gene A′ and B′ in *Caenorhabditis elegans* are orthologs of gene A and B in *Saccharomyces cerevisiae*. Gene A′ and B′ in *S. cerevisiae* are paralogs of gene A and B in *S. cerevisiae*. Co-expression is defined by a correlation in the expression profile higher than 0.6, indicated by the arrow. We define a gene pair A-B to be a duplicated pair with conserved co-expression when the genes are co-expressed and their closest relatives (lowest, significant E-value in Smith–Waterman searches) A′ and B′ are also co-expressed. Orthologous conservation is the conservation of co-expression between A–B and A′–B′ between two species.

Conserved co-expression improves accuracy of pathway prediction

Does the conservation of co-expression after gene duplication or speciation increase the likelihood of a functional relationship between co-expressed genes? Conservation after duplication in *S. cerevisiae* does indeed increase the accuracy levels for prediction of functional interactions, albeit at the expense of coverage of known interactions (Fig. 2.1). The results for *C. elegans* are similar, but there are not enough genes annotated in the PATHWAY database to establish the accuracy for conserved co-expression above 0.6. Higher accuracy is also achieved for the genes that are co-expressed in both species (Fig. 2.1). A similar result was described by Teichmann and co-workers (Teichmann and Babu, 2002), who found that 89% of the conserved co-expressed pairs between *S. cerevisiae* and *C. elegans* for which functional annotation was available were part of the same protein complex. However, in this analysis co-expression was defined in such a strict way that 93% of the conserved pairs already had a functional annotation and hardly any new predictions could be made. Note that our orthology prediction based on phylogenetic trees, instead of the Bidirectional Best Hit (Overbeek et al., 1999) method, allows ~50% more predictions to be made at a correlation higher than 0.6: instead of 799 there are 1215 predicted interactions in *S. cerevisiae*, and instead of 607 there are 803 predicted interactions in *C. elegans*.

Predicted interactions of *S. cerevisiae* genes were verified not only by the PATHWAY database, but also by using gene ontology (GO) annotations (Ashburner et al., 2000; Dwight et al., 2002). When involvement in the same biological process is defined as a common GO process category at the fourth level of specification, the accuracy achieved at a correlation threshold of 0.6 is 93% using orthologous conservation and 82% using paralogous conservation, compared with only 31% for all co-expressed pairs. There are insufficient reliable GO annotations on *C. elegans* genes (most are inferred by electronic annotation) to confirm their predicted interactions using GO.

Conserved co-expressed gene pairs for which only one of the genes is assigned to a pathway form a pool of genes to which we can now assign a pathway. From interspecies or intraspecies conservation, we predict a pathway for 55 and 95 *S. cerevisiae* genes, and for 54 and 596 *C. elegans* genes, respectively. For the vast majority of genes found by paralogous conservation (282 in *S. cerevisiae*, 2216 in *C. elegans*) and by orthologous conservation (91 in *S. cerevisiae*, 143 in *C. elegans*), neither gene in the pair is present in the PATHWAY database.

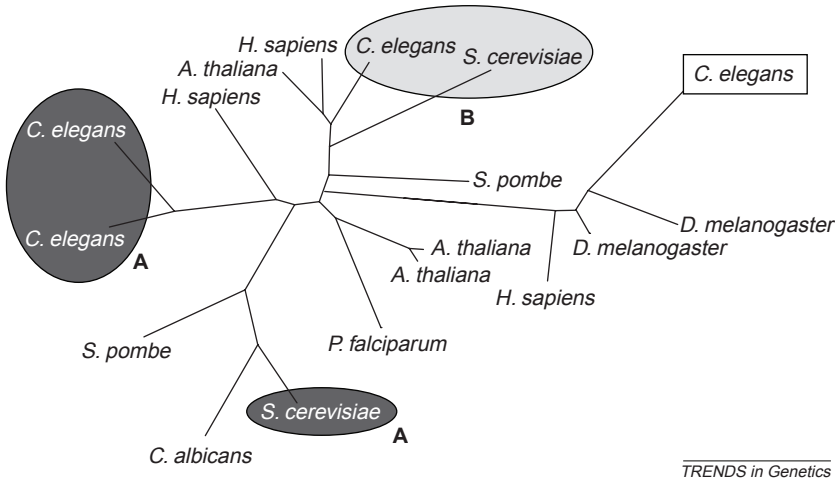


Figure 2.3 | Orthology prediction using an unrooted phylogenetic tree. Large-scale orthology prediction is generally done by the Best Bi-directional Hit approach or extensions thereof like COGs (Tatusov et al., 2001). As orthology is an evolutionary relation we determine it using phylogenetic trees. Our method includes also inparalogs and is conceptually similar to INPARANOID (Remm et al., 2001). All predicted protein-coding genes were obtained for both *Saccharomyces cerevisiae* (Goffeau et al., 1996) and *Caenorhabditis elegans* (1998), as well as predicted genes of other complete genomes (to improve the quality of calculated phylogenies). Each *S. cerevisiae* gene is considered in turn to find orthologs in *C. elegans*. First we find homologies between all predicted genes and the gene under consideration by Smith–Waterman searches(Pearson, 1998; Smith and Waterman, 1981). We include all genes with an E-value smaller than 0.01 and of which the region of homology is larger than the 50% of the length of the query. Groups of more than 250 proteins are reduced in size by applying a lower E-value cutoff. A multiple alignment is made with ClustalW (Thompson et al., 1994) from the protein sequences of the gene and its homologs and a Neighbor-Joining (Saitou and Nei, 1987) tree is calculated. For every query gene, we first select the largest branch containing the query gene and possible paralogs in *S. cerevisiae*, but no *C. elegans* genes. And after that the smallest branch that contains this branch as well as *C. elegans* genes, but no extra *S. cerevisiae* genes is selected. Orthology is assigned between all *S. cerevisiae–C. elegans* pairs in this branch. This results in the assignment of orthologous relations to the genes in the dark grey circles, indicated with A, and to the genes in the light grey circle, indicated with B. The boxed *C. elegans* gene has no orthologs in *S. cerevisiae*.

New predictions from old data

Co-expression conserved between S. cerevisiae and C. elegans of the hypothetical gene CAT5 (YOR125C, ZC395.2) and COQ2 (YNR041C, F57B9.4) confirms earlier predictions based on knock-out experiments (Marbois and Clarke, 1996) and homology relations (Rea, 2001) that CAT5 is 2-polyprenyl-3-methyl-6-methoxy-1,4-benzoquinone mono-oxygenase, which is involved in ubiquinone synthesis, as COQ2 encodes para-hydroxybenzoate: polyprenyl transferase, which is also involved in ubiquinone synthesis.

A prediction based on conservation of co-expression after duplication concerns the link between YBR052C and YDR074W. The gene YBR052C probably catalyzes a redox reaction, because it belongs to the WrbA family, which is homologous to flavodoxins. The gene YDR074W encodes trehalose-6-phosphatase (De Virgilio et al., 1993), which is involved in starch and sucrose metabolism. For one redox enzyme in this pathway, glucoside 3-dehydrogenase, no gene has been described yet. This enzyme binds the co-factor flavin mononucleotide (FMN) (Hayano and Fukui, 1967) and has a molecular mass of 85 kDa (van Beeumen and de Ley, 1975) in Agrobacterium tumefaciens, where an ortholog of YBR052C is also present. The Escherichia coli ortholog, WrbA, forms multimers and also binds FMN (Grandori et al., 1998). The amino acid sequence of WrbA indicates a molecular mass of 22

kDa, implying a tetrameric organization consistent with the formation of multimers and the determined molecular mass of 85 kDa. We thus propose that YBR052C encodes the enzyme glucoside 3-dehydrogenase.

A more speculative prediction is that YKL033W-A (R151.8), whose co-expression with the endonuclease APN1 (T05H10.2) is conserved between species, is a 3' phosphatase involved in DNA repair. The gene YKL033W-A contains a frameshift in the sequence of the published S. cerevisiae genome, but has also been sequenced without a frameshift (Purnelle et al., 1994) (accession number X71622) and has full-length orthologs in all sequenced eukaryotes. The human ortholog, GS1, is particularly interesting as it is an X-chromosome gene that escapes X inactivation (Yen et al., 1992). The protein is homologous to haloacid dehalogenase-like hydrolases, a domain that has phosphatase activity, and is among others found as a 3' phosphatase in T4 tRNA-repair enzyme, polynucleotide kinase (Galburt et al., 2002). DNA 3' phosphatase reactions do have a role in repairing lesions in the DNA. This process involves APN1, which exhibits 3' phosphodiesterase activity (Vance and Wilson, 2001).

Modularity in pathway evolution

Of particular evolutionary importance is our finding of a substantial number of cases where, although the expression pattern of A' and B' has changed relative to their ancestors A and B, the co-expression of A' and B' is conserved. This seemingly contradicts the finding by Wagner that after duplication events, mRNA expression patterns diverge very quickly relative to amino acid sequence (Wagner, 2000). Yet, both results complement each other as we show that the co-expression is often conserved even when the expression patterns are not. However, a real contradiction with our results is apparent in a study of small molecule metabolism pathways in E. coli that showed that modular recruitment occurs very rarely (Teichmann et al., 2001). Our observation of co-duplicated, diverged but still co-expressed genes suggests a substantial role for modularity in pathway evolution.

Outlook

Correlations between expression profiles do not necessarily imply co-regulation, and co-regulation does not always indicate functional interaction. Thus, it is important for function prediction to increase the reliability of co-expression data. Overlapping transcriptional clusters from different clustering methods have led to the prediction of functional categories for many genes (Wu et al., 2002). Here we show that both intraspecies and interspecies conservation make expression data useful for the reliable prediction of specific functions.

Both types of conservation differ in their future applicability. Paralogous co-expression conservation has great advantages, because it relies on experimentation in only a single organism. Moreover, gene duplications are rampant in eukaryotes. The resulting noise in orthology prediction possibly distorts the usage of conservation of co-expression between species. However it is the very same gene duplication that increases the applicability of co-expression for function prediction.

Acknowledgements

This work was supported in part by a grant from the Netherlands Organization for Scientific Research (NWO).

Chapter 3

The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model

Vera van Noort, Berend Snel, and Martijn A Huynen

EMBO Reports 5, 280-284 (2004)

The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model

Abstract

We investigated the gene coexpression network in *Saccharomyces cerevisiae*, in which genes are linked when they are coregulated. This network is shown to have a scale-free, small-world architecture. Such architecture is typical of biological networks in which the nodes are connected when they are involved in the same biological process. Current models for the evolution of intracellular networks do not adequately reproduce the features that we observe in the network. We therefore derive a new model for its evolution based on the observation that there is a positive correlation between the sequence similarity of paralogues and their probability of coexpression or sharing of transcription factor binding sites (TFBSs). The simple, neutralist's model consists of (1) coduplication of genes with their TFBSs, (2) deletion and duplication of individual TFBSs and (3) gene loss. A network is constructed by connecting genes that share multiple TFBSs. Our model reproduces the scale-free, small-world architecture of the coregulation network and the homology relations between coregulated genes without the need for selection either at the level of the network structure or at the level of gene regulation.

Introduction

Unravelling the interactions between the elements of a cell constitutes a major goal of the genome era. The structure of the resulting interaction networks is relevant to the functioning of the cell, for example, in development (Davidson et al., 2003), and for the interpretation of experimental results. Network analyses have shown a correlation between, on the one hand, the essentiality of a gene and, on the other hand, either the number of connections that the gene has (Jeong et al., 2001) or the topology of the metabolic network (Forster et al., 2003; Stelling et al., 2002). Furthermore, networks provide, for example, a framework for the interpretation of synthetic lethal knockouts (Brummelkamp and Bernards, 2003; Sonoda et al., 2003). The analysis of intracellular network topology also provides an objective, genome-wide base for the classic idea that a cell can be divided into functional modules (Davidson et al., 2003; Snel et al., 2002; Yanai and DeLisi, 2002), and network topology correlates with sequence variation: sequences evolve slowly when they have many connections in the network (Fraser et al., 2002) or when they are part of relatively densely connected motifs (Wuchty et al., 2003). Finally, network approaches are used to integrate various types of genomics data to increase the reliability of predicted interactions (Jansen et al., 2003), and one can envision that the topology of intracellular networks provides constraints for the manipulation and design of cells.

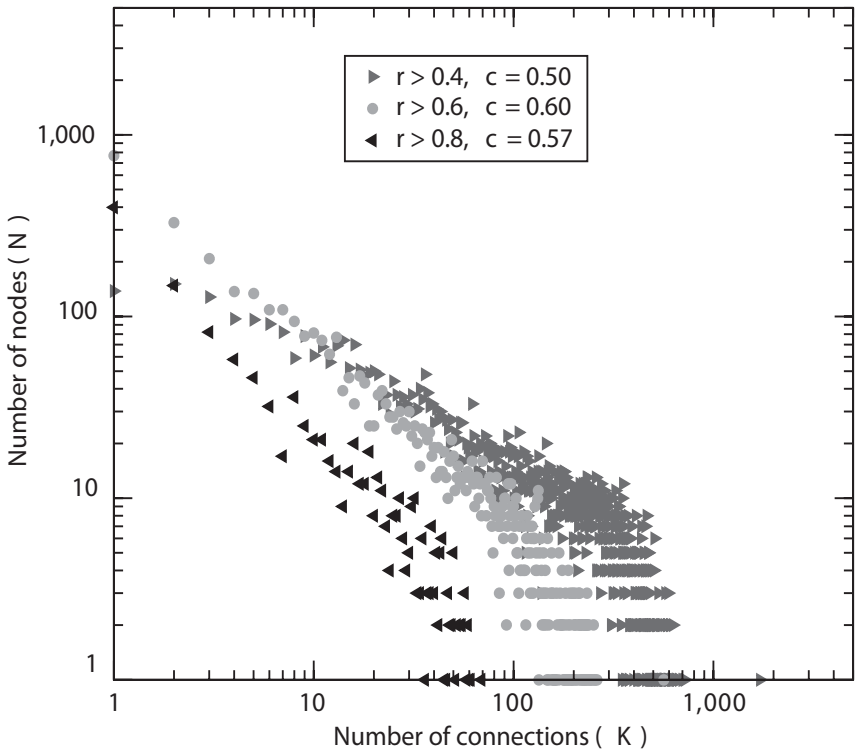


Figure 3.1 | Distribution of connections per node in the coexpression network. Nodes are genes and connections are defined by coexpression of two genes, resulting in a network. The number of nodes (N) with a certain number of connections (k) in the coexpression network is shown, where coexpression is defined by a correlation in expression pattern higher than 0.4 (right-pointing arrows), 0.6 (circles) or 0.8 (left-pointing arrows). The distributions at thresholds 0.6 and 0.8 are scale free with an exponent $\gamma \approx 1$.

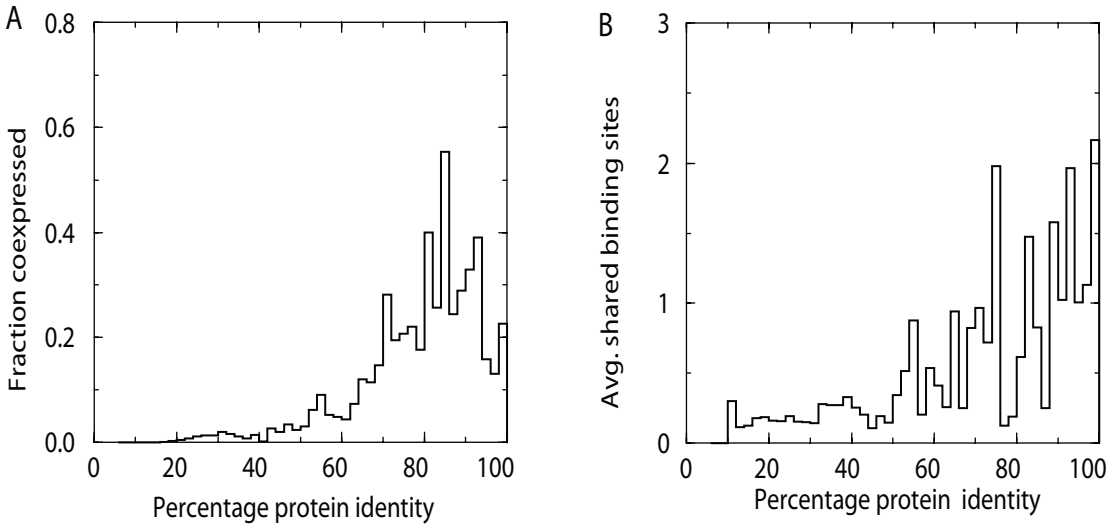


Figure 3.2 | Coexpression between paralogues in experiments. (A) Fractions of coexpressed paralogues calculated by correlation in coexpression in the data set of (Hughes et al., 2000). (B) Average number of shared regulatory elements between paralogues in the data set of (Lee et al., 2002).

The main source of data for the reconstruction of intracellular networks is genomics. Facets of the cellular network that have been studied include protein interaction networks in which the nodes (proteins) are connected when they physically interact (Ito et al., 2001; Jeong et al., 2001; Uetz et al., 2000; Wagner, 2001), metabolic networks in which the nodes (metabolites) are connected when they are substrates or products in the same biochemical reaction (Fell and Wagner, 2000; Jeong et al., 2000; Ma and Zeng, 2003), genomic association networks in which the nodes (genes) are connected when they occur repeatedly together in operons (Snel et al., 2002), and evolutionarily conserved coexpression networks (Stuart et al., 2003). The study of these networks has revealed that they all have a similar, nontrivial architecture. First, they are so-called scale-free networks. This means that there is no typical number of connections per node; rather the distribution of the number of connections (k) per node (N) follows a power law ($N(k) \sim k^{-\gamma}$). In other words, there are many nodes with few connections and a small but still significant number of nodes with many interactions. Second, these networks have a small-world architecture. This implies that, on the one hand, they are highly clustered: when a node is connected to two other nodes, the latter two also tend to have a direct connection to each other. On the other hand, the average shortest path length in the network (L, the minimum number of connections that one needs to get from one node to any other node) is almost as low as that for random networks (Watts and Strogatz, 1998). The scale-free, small-world architecture appears typical for intracellular networks in which the nodes are connected when they are involved in the same biological process. In contrast, another type of network, the gene regulatory network of *Saccharomyces cerevisiae*, in which the connections are between transcription factors and the genes they regulate, does not have a scale-free but rather an exponential distribution of the number of connections per node (Guelzim et al., 2002; Lee et al., 2002).

Because of the importance of molecular networks for the functioning of the cell, there is a great deal of interest in the evolution and origin of these networks. Yet it remains an open question whether the scale-free, small-world architecture is a direct product of selection and thus functionally meaningful, merely a by-product of the requirements of function and of selection at other levels, or even a natural consequence of mechanisms such as gene duplication. The evolution of scale-free networks has been explained in terms of selection on global properties such as robustness (Guelzim et al., 2002; Jeong et al., 2000) and the fast spread of perturbations (Fell and Wagner, 2000). It has also been addressed in phenomenological models (Bhan et al., 2002; Ravasz et al., 2002) that do not require selection but that are not supported by independent data. Here we analyse the network architecture of a general indicator of protein involvement in the same biological process: gene coexpression in *S. cerevisiae* (Hughes et al., 2000). We show that the gene coexpression network in *S. cerevisiae* is a scale-free, small-world network. By exploiting homology relations between the genes in the coexpression network, we formulate a neutralist model in which the scale-free, small-world architecture is a natural consequence of the mechanisms behind gene regulation evolution. This calls into question global selection mechanisms for the architecture of intracellular networks.

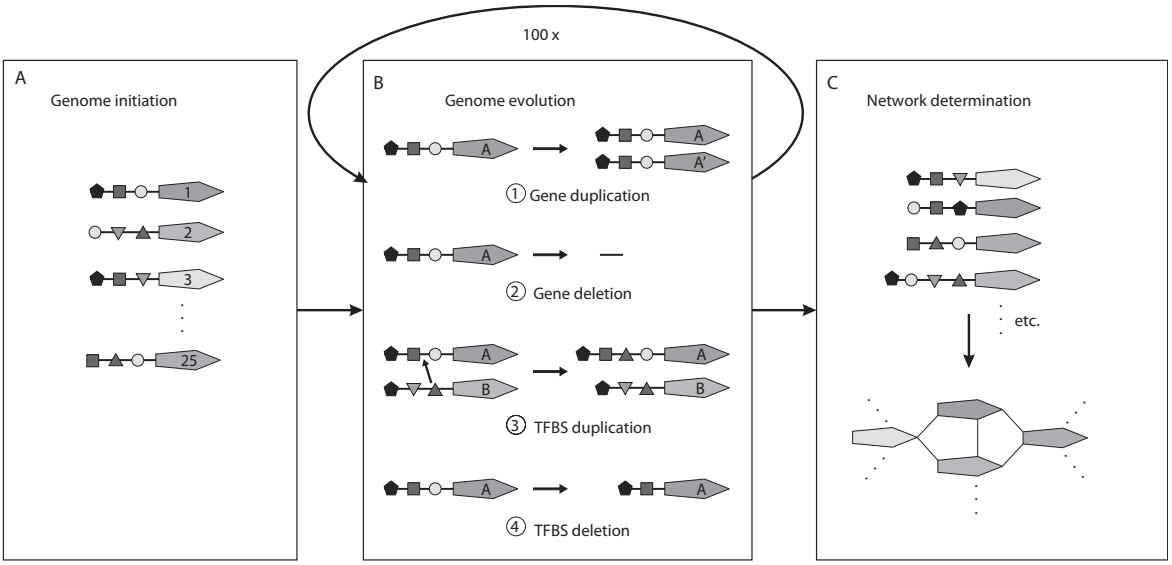


Figure 3.3 | Evolutionary model of transcription regulation. **The evolutionary model consists of a few simple mechanisms. (A) A genome is initiated with 25 genes with random TFBSs, represented by the small coloured shapes. (B) Possible events are as follows: (1) Gene A is duplicated, gene A0 has the same TFBS as its duplicate gene A; the duplicates are coexpressed. (2) Gene deletion. (3) Gene A acquires a new TFBS from gene B. The probability of obtaining a specific TFBS is proportional to its frequency in the genome. The probability of a novel TFBS is (150 - total number of different TFBSs present) / (150 + total number of TFBSs). (4) One of the TFBSs of gene A is deleted. (C) A network is constructed by connecting genes that share TFBSs.**

Results

Although gene coexpression is a continuous observable, the underlying principle is discrete: the sharing of regulatory elements. We therefore translate gene coexpression into a discrete network. In the network, the genes are the nodes, which are connected to each other when coexpressed. Such a network representation allows a comparison of the global organization of gene expression with other facets of the intracellular network. Furthermore, relative to protein interaction networks or metabolic networks, coexpression covers a more inclusive array of functional relations between gene products. As a threshold to establish a link in the network between two genes, we chose a coexpression correlation of 0.6 in a large-scale expression data set (Hughes et al., 2000), as higher thresholds do not give higher reliabilities of functional interaction between the encoded proteins (van Noort et al., 2003). The coexpression network has 4,077 nodes (genes) that are linked by a total of 65,430 connections, the average number of connections per node (k) thus being 32 (each connection links two nodes). The distribution of number of links per node is scale free with degree exponent $\gamma \approx 1$ (Fig 3.1). Note that although the average number of connections is 32, most genes are connected to only one other gene, as reflected by the scale-free distribution (Fig 3.1). The clustering coefficient of the network (c, the fraction of cases where if a node has a connection to two other nodes, these two also have a direct connection to each other) is 0.6. Not all nodes are connected in one cluster; the largest cluster contains 3,945 nodes, with an average shortest path length (L) of 4. In a random network with the same number of nodes (N) and connections (k), $c=0.008$ (k/N) (Barabasi and Albert, 1999) and $L \approx 2.8$ (from simulations; see Methods). Thus, the yeast coexpression network has all the properties of a small-world

($L \approx L_{random}$, $c >> c_{random}$), scale-free ($N(k) \sim k^{-\gamma}$) network that is typical for intracellular networks in which the nodes are connected when they are involved in the same process. Using thresholds for coexpression higher than a correlation coefficient of 0.6 gave similar results, that is, a scale-free degree distribution and small-world organization (Fig 3.1). Using lower thresholds leads to the inclusion of 'random' connections (van Noort et al., 2003) and an exponential degree distribution with a smaller c (Fig 3.1). At the threshold of 0.6, the network statistics are similar to previously studied biological networks (Fell and Wagner, 2000; Jeong et al., 2001; Jeong et al., 2000; Snel et al., 2002; Wagner, 2001), and thus we use this network for further study.

The coexpression data have another interesting property: a correlation between the fraction of coexpressed paralogues and their sequence similarity (Fig 3.2A). An independent data set that also contains this pattern is the large-scale, experimental determination of transcription factor binding sites (TFBSs) (Lee et al., 2002), in which the number of shared regulatory elements between paralogues increases with protein identity (Fig 3.2B). A correlation between divergence in sequence and in coexpression is expected if both diverge at constant, clock-like rates (Wagner, 2000), and indicates neutral evolution of these two traits. It appears that in the case of gene duplication, the regulatory elements tend to be coduplicated with the genes and mutated afterwards.

Existing network-evolution models cannot account for the combination of the architecture of the coexpression network and the correlation between coexpression and sequence similarity in paralogues. The network model of (Barabasi and Albert, 1999), based on the concept of preferential attachment (Simon and Bonini, 1958), produces scale-free networks, but not small-world networks ($c \approx c_{random}$; in a small-world network $c >> c_{random}$), even when

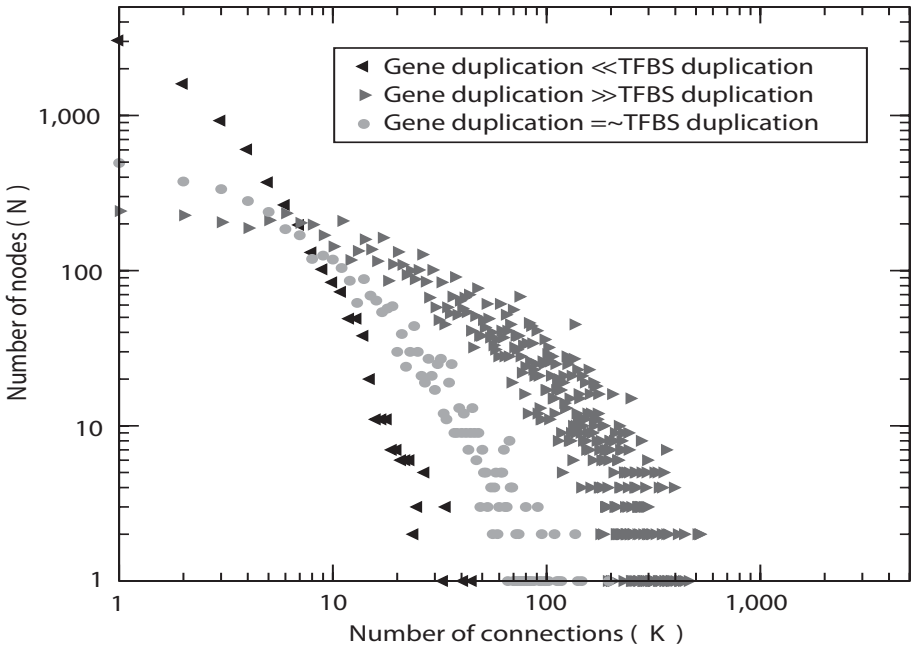


Figure 3.4 | Distribution of connections per node in the simulated network. The number of nodes (N) with a certain number of connections (k) in the simulated network is shown. The minimum number of shared TFBSs for a connection in the network is three. Gene duplication and deletion are in the same order of magnitude as TFBS duplication and deletion (circles), gene duplication and deletion are much smaller than TFBS duplication and deletion (left-pointing arrows), and gene duplication and deletion are much larger than TFBS duplication and deletion (right-pointing arrows).

introducing constraints to the number of connections per node or to the ageing of nodes (Amaral et al., 2000). The algorithm of (Ravasz et al., 2002) to realize a small-world, scale-free network involves hierarchical duplication of complete modules and attachment to the central node of the existing module. This model does not lead to a high likelihood of attachment between duplicated nodes, and is therefore not explanatory for the evolution of our network. Moreover, in contrast to the predictions of this model, the explicit testing of the age of genes (see Methods) and the number of their connections did not reveal any positive correlation (Pearson correlation=–0.04, P-value that there is no positive correlation=0.98). The duplication model of (Bhan et al., 2002) assumes duplication of genes with partial conservation of connections. When seeding this model with a scale-free network, most of the structure persists for a few iterations; however, simulating this model for a higher number of iterations results in an exponential degree distribution of N versus k (Pastor-Satorras et al., 2003). In this model, there is no relation between the timing of a duplication event and the likelihood of attachment of the resulting paralogues. This is because the connections are fixed once established, as in all previous models. This is not an evolutionarily sound assumption, given the observation that connectivity between paralogues is dependent on the timing of the duplication event and that coexpression is only partly conserved between species (Teichmann and Babu, 2002; van Noort et al., 2003).

We introduce a new, simple model to explain the emergence of scale-free networks with a high clustering coefficient that is based on the observation of a positive correlation between the probability of a connection between two paralogues and their sequence similarity. In this model, the entities are genes that have a number of TFBSs. Connections between genes are established when they share a minimum number of TFBSs. At every time step, each gene has a probability of being duplicated, resulting in a new gene (step 1, Fig 3.3). In the case of duplication, the TFBSs are passed on to the duplicate gene, corresponding to a high likelihood of coexpression between recently duplicated paralogues in the experimental data. A gene may be deleted (step 2, Fig 3.3). A TFBS can be acquired from the pool of TFBSs of all genes, where the probability of obtaining a specific TFBS is proportional to its frequency in the genome (step 3, Fig 3.3), introducing connections between nonparalogous genes. New TFBSs are introduced at a low frequency. All TFBSs have a probability of being deleted (step 4, Fig 3.3), giving rise to a decrease in connectivity between duplicates over time and balancing the number of TFBSs per gene. We simulated this model by seeding it with 25 genes with randomly assigned TFBSs and evolving these for 100 evolutionary steps, observing three parameter regimes. In the first regime (left-pointing arrows, Fig 3.4) , the TFBS duplication and deletion rates are much higher than the gene rates. This effectively decouples the TFBS from the genes and gives rise to a very loosely connected network (a steep slope), albeit with a power-law distribution of the number of connections per node and a high c (c=0.3 in this specific case). In the second regime (circles, Fig 3.4), the TFBS duplication and deletion rates are in the same order of magnitude as those for the genes. Here, we observe a scale-free degree distribution with a slope similar to the one observed in the experimental data and a high c. In the third regime (right-pointing arrows, Fig 3.4), the rates for TFBS duplication and deletion rates are much lower than those for genes. This couples the TFBS to the genes such that almost every pair of paralogues is connected, resulting in a very tightly

connected network, with an exponentially declining degree distribution and a very high c (close to 1).

In a natural situation, we do not expect the evolutionary parameters to be in the third regime, as pieces of DNA are duplicated by the same mechanisms, be it coding or noncoding DNA. Also, TFBSs are much smaller than genes and are thus expected rather to have duplication and deletion rates that are at least as high as those for individual genes. A simulated network in the intermediary regime exists of, for example, 4,273 nodes connected by 56,953 connections. The network displays small-world behaviour, indicated by a high clustering coefficient ($c=0.2$) relative to random networks ($c_{random}=0.003$) and in the largest cluster of 4,070 nodes an average shortest path length ($L\approx3$) that is similar to the shortest path length in a random network ($L_{random}\approx3.5$). The overall behaviour of this network is very similar to the coexpression network. This indicates that a scale-free, small-world organization as such can be the result of neutral evolution. Still, the levels of cliquishness and the slope of the scale-free distribution may be the result of natural selection.

Discussion

The functional relevance of the typical scale-free, small-world organization that we observe in intracellular networks is open to debate. In the absence of an experimental system with which to test the functional relevance of the network architecture, we have to resort to theoretical experiments. These basically answer the following question: what are the minimal conditions under which a specific network architecture can evolve? To answer these questions, we have studied the coexpression network in *S. cerevisiae* that we show to have a small-world, scale-free architecture. Furthermore, the network contains a positive correlation between the probability of coexpression of two paralogues and their sequence similarity. We introduce a network model that reproduces the architecture as well as the homology relations in the coexpression network. Its key components are that genes are coduplicated with their TFBSs and that multiple shared TFBSs are required for coexpression. Our observation of a positive correlation between sequence similarity and the level of coexpression contrasts with the results of (Wagner, 2000), who only observed a very weak correlation. The difference is probably explained by the much larger coexpression data (Hughes et al., 2000) and the additional data set of TFBSs (Lee et al., 2002) combined with homology relations. This analysis of more data thus offers support for a neutralist's explanation of the gene coexpression network architecture.

In contrast, not only the scale-free, small-world architecture of intracellular networks but also one of the network statistics, the diameter, have been argued to be the result of biological selection. It should be noted that with respect to the diameter, the direction of this argument has been rather arbitrary: both the relatively small diameter of metabolic networks (Jeong et al., 2000) and the relatively large diameter of protein interaction networks (Maslov and Sneppen, 2002) have been argued to be the result of selection. Subsequent analyses have however shown that in both cases the networks were more random than proposed, and that the observed biases in the diameter size were either due to the choice of the network nodes (Ma and Zeng, 2003) or experimental bias in the underlying data set

(Aloy and Russell, 2002). This leaves the argument that the scale-free, small-world architecture itself is a result of selection (Guelzim et al., 2002). As our model is purely mechanistic and the mechanisms are sufficient to explain the properties of the network, we do not need selection at the level of the network or at the level of gene regulation. This does not exclude the possibility of selection at that level or that the network architecture is in some way or another exploited by the cell, but it does call for a more sober view in interpreting network architectures in terms of selection and the benefits for the cell.

Methods

Random network. To evaluate the nontrivial properties of the coexpression network, it is compared with a random network. The random network is simulated by taking the same number of nodes as the coexpression network and randomly placing the same number of connections between these nodes.

Clustering coefficient and average shortest path length. The clustering coefficient (c) or the degree of cliquishness is computed by first counting all pairs of associations (cases where gene A is linked to gene B and to gene C), subsequently counting how often these pairs are closed (B is linked to C), and then dividing the second count by the first count (Watts and Strogatz, 1998). L is the average minimum number of nodes one needs to cross to get from one node to another. To obtain L, we compute the shortest path between all pairs of genes, and subsequently compute the average (Watts and Strogatz, 1998).

Gene age. The age of genes was determined by the amino-acid distance (100 – percentage protein identity) to the most distant paralogue (homologue within the same genome; (Fitch, 1970)). Duplications seem to be rampant in yeast; thus, when a gene was present very early in the genome, it is likely to have distant paralogues. This distance was then used to find out whether there is a correlation between gene age and the number of connections in the coexpression network.

Paralogues. To determine the correlation between protein identity and probability of connections between paralogues, we first need to determine paralogues. This is done by Smith–Waterman (Smith and Waterman, 1981) searches of the amino-acid sequences of the translated genes of *S. cerevisiae* (Goffeau et al., 1996) against each other. Matches with an E-value below 0.01 are considered paralogues.

Acknowledgements

This work was supported in part by a grant from the Netherlands Organization for Scientific Research (NWO).

Chapter 4

Comparative genomics for reliable protein-function prediction from genomic data

Martijn A Huynen, Berend Snel and Vera van Noort

Trends in Genetics 20, 340-344 (2004)

Comparative genomics for reliable protein-function prediction from genomic data

Abstract

Genomic data provide invaluable, yet unreliable information about protein function. However, if the overlap in information among various genomic datasets is taken into account, one observes an increase in the reliability of the protein-function predictions that can be made. Recently published approaches achieved this either by comparing the same type of data from multiple species (horizontal comparative genomics) or by using subtle, Bayesian methods to compare different types of genomic data from a single species (vertical comparative genomics). In this article, we discuss these methods, illustrating horizontal comparative genomics by comparing yeast two-hybrid (Y2H) data from *Saccharomyces cerevisiae* with Y2H data from *Drosophila melanogaster*, and illustrating vertical comparative genomics by comparing RNA expression data with proteomic data from *Plasmodium falciparum*.

Introduction

Functional genomics data, derived from proteomics and transcriptomics, enable us to have an unprecedented view of global cellular activity. However, these data are 'noisy': they miss many of the true protein interactions and they also report numerous protein interactions that are false. Fortunately, computational analysis of these data can improve our ability to extract reliable predictions from them. Horizontal comparative genomics achieves this by comparing multiple datasets of the same type that are derived from different species. It thus compares not only two independent 'human' experiments but also experiments performed by evolution. It can help answer whether genomic data indicate that proteins from two orthologous groups functionally interact with each other in multiple species.

A classic example of horizontal comparative genomics used the conservation of gene order in prokaryotes as an indication of the co-regulation of proteins in multiple species (Dandekar et al., 1998). This principle has now been applied to gene co-expression data that were determined from array data and to protein–protein interaction data that were obtained through Y2H screens. The likelihood that the observed links between the proteins are biologically meaningful increases dramatically when gene co-expression between orthologous mRNAs is conserved between *Saccharomyces cerevisiae* and *Caernohabditis elegans* (Teichmann and Babu, 2002; van Noort et al., 2003) or among *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *Homo sapiens* (Stuart et al., 2003) or when Y2H interactions are conserved between Helicobacter pylori and *S. cerevisiae* (Kelley et al., 2003). This likelihood is measured as the fraction of proteins (among those co-expressed or Y2H interacting proteins whose functions are known) that are part of the same complex, pathway or biological process.

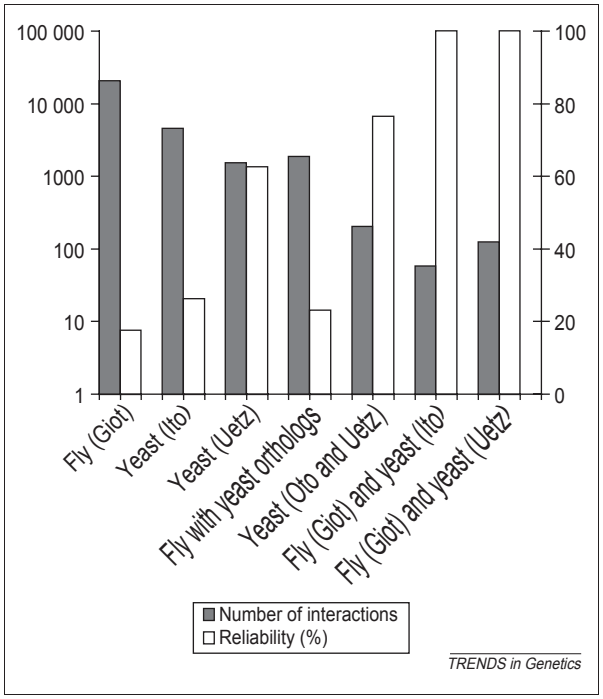


Figure 4.1 | The reliability of separate and combined yeast two-hybrid (Y2H) data for the prediction of functional interactions between proteins (left), and the total number of interactions detected (right). Reliability is measured as the number of proteins for which Y2H interactions are observed (on the same Kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa et al., 2004) pathway-map) divided by the total number of proteins for which Y2H interactions are observed (in the KEGG database; http://www.genome.ad.jp/kegg). The three columns on the left-hand side contain data from the Drosophila melanogaster dataset of (Giot et al., 2003), the Saccharomyces cerevisiae datasets from (Ito et al., 2001) and from (Uetz et al., 2000). In the center are the columns that give the reliability and the number of Y2H interactions that were measured in D. melanogaster for which both proteins have orthologs in the S. cerevisiae genome. The columns on the right show the reliability for the Y2H interactions that are observed in two datasets. Orthology was defined using the Clusters of Orthologous Groups (COG) database. Conservation of Y2H interaction between S. cerevisiae and D. melanogaster leads to a greater increase in reliability than the independent observation of that interaction in the two datasets from S. cerevisiae, and the fact that two Y2H interacting D. melanogaster proteins both have an ortholog in S. cerevisiae has little effect on the reliability of the interaction. Note, however, that the reliability of conserved Y2H interactions does come at the price of sensitivity: <200 interactions are conserved between the combined S. cerevisiae and D. melanogaster datasets. Using the more restrictive best bi-directional hits, the overlap is <100 (Table 4.1). For complete lists of the conserved interactions and the effects of using different orthology definitions see http://www.cmbi.kun.nl/~huynen/ConservedY2H.

Interestingly, another type of evolutionary conservation, conservation of co-expression or Y2H interaction after parallel gene duplication in one species, leads to a similar increase in the reliability of the predictions that can be made (Kelley et al., 2003; van Noort et al., 2003). Conservation of co-expression can be used to predict protein function and functional interactions reliably (Stuart et al., 2003; van Noort et al., 2003) and some predictions have been verified experimentally (van Noort et al., 2003).

Function prediction by conserved interaction

There are some conceptual and technical issues that are involved in comparing genomic data from different species, which we illustrate by comparing the recently published Y2H data from D. melanogaster (Giot et al., 2003) with Y2H data from S. cerevisiae (Ito et al., 2001; Uetz et al., 2000). Similar to observations made in other horizontal comparative genomics analyses, we found that detecting a Y2H interaction between two orthologous groups in two species dramatically increases the likelihood that they interact functionally (Figure 4.1); an inspection of the list of conserved interactions indicates that they are all physical interactions. The total number of conserved interactions is however rather low (Figure 4.1), indicating the high reliability and low sensitivity of using the conservation of Y2H interactions to predict physical interactions between proteins reliably. Furthermore, based on the Saccharomyces Genome Database (SGD) (http://www.yeastgenome.org/) (Dwight et al., 2002), the list of conserved Y2H interactions contains a few proteins for which the biological processes and the molecular functions are unknown (13 proteins, 5% of the proteins in conserved in-

teractions), compared with the complete genome (27% genes with unknown function). This parallels the observation in the *S. cerevisiae* interaction network, where proteins of known function have more interaction partners than those of unknown function (Yu et al., 2004).

Thus conserved interactions often give credence to cases for which there is already at least some experimental evidence, such as the interaction of the *S. cerevisiae* protein TSR2 and its ortholog in *D. melanogaster* (CG14543) with ribosomal protein 26S in either species. A role for YLR435w in ribosomal maturation would be consistent with the accumulation of 20S rRNA in YLR435w knockouts in *S. cerevisiae* (Peng et al., 2003) and an interaction with the ribosomal protein 26S in *S. cerevisiae* has also been observed using tandem affinity purification (TAP)-tagging (Peng et al., 2003).

Nevertheless, using overlapping datasets, one can make new, reliable protein–protein interaction predictions. One example of a 'new' interaction is between the xeroderma pigmentosum group A binding GTPase (XAB1) (CG3704 in *D. melanogaster*) and the hypothetical protein YOR262w/CG10222, which also contains a GTPase domain. XAB1 has been observed to interact with the DNA-repair protein XPA1 and is thought to be required for its import into the nucleus (Nitta et al., 2000), suggesting a function in nuclear import for YOR262w/CG10222. The two proteins share other, albeit weak, genomic links: in *S. cerevisiae,* both proteins are cytoplasmatic (Huh et al., 2003) and are essential for the cell (Dwight et al., 2002). Furthermore, they have the same phylogenetic distribution, possessing orthologs in all eukaryotic genomes sequenced to date.

More than just more data

Is combining data from different species really more than just combining multiple independently generated datasets to filter out the experimental noise? There are some indications that this is the case: Stuart and coworkers showed that omitting parts of the co-expression data in each species did not significantly affect their ability to reliably predict the proteins that were part of the same pathway (Stuart et al., 2003). By contrast, omitting one or more species from the data drastically reduced this predictive value (Stuart et al., 2003). Furthermore, although taking the overlap into account between two different Y2H datasets for *S. cerevisiae* also leads to an increase in the likelihood of detecting a real interaction, it does not match the level that is obtained by comparing datasets from multiple species (Figure 4.1). Note however that when two genes are co-expressed in *S. cerevisiae* the fact that they both have orthologs in *C. elegans* in itself already leads to a higher likelihood of interaction, although not as high a likelihood as when these orthologs are also co-expressed in *C. elegans* (van Noort et al., 2003). Thus, the (same) widespread, phylogenetic distribution of genes appears to be responsible in part for the increase in the predictive value of finding conserved co-expression between them. This effect is however not observed for the Y2H data (Figure 4.1), where the presence of orthologs in the other species does not significantly affect the reliability of the interaction.

Quantifying the amount of conservation and determining orthology

The amount of conservation of co-expression between *S. cerevisiae* and *C. elegans* is low (<10%) (Teichmann and Babu, 2002; van Noort et al., 2003), albeit significant (van Noort et al., 2003). It is not clear to what extent the small overlap is a reflection of the noisy nature

Table 4.1 | The overlap among the number of yeast two-hybrid interactions between two Saccharomyces cerevisiae datasets and one Drosophila melanogaster dataset.[a]

Dataset comparison	Protein interactions (both proteins present in the other dataset)	Conserved interactions	Percentage of conserved interactions	Mean conserved interactions
Ito versus Uetz	858; 697	201	23.4%; 28.8%	26.1%
Ito versus Giot	229; 394	45	19.6%; 11.4%	15.5%
Uetz versus Giot	120; 168	33	27.5%; 11.6%	23.5%

[a]In calculating the percentage of overlap between datasets only interactions for proteins that appeared in both datasets were taken into account (i.e. the number of interactions that was observed in both S. cerevisiae datasets was divided by the number of interactions in the Ito set, for which both proteins were present, although not necessarily interacting with each other, in the Uetz set) Orthology relations between the S. cerevisiae and the D. melanogaster were determined by best bidirectional hits among the homologous relationships (E , 0.01, local sequence alignment). The percentage overlap between the S. cerevisiae datasets and the D. melanogaster dataset (24% and 16%) are close to those between the S. cerevisiae datasets themselves (26%). This suggests that the low levels (24% and 16%) of conservation of physical interaction between S. cerevisiae and D. melanogaster can, to a large extent, be attributed to the lowreproducibility of yeast two-hybrid (Y2H) interactions in general; therefore, physical interaction between proteins is highly conserved in evolution.

of the data or a true indication of the low conservation of co-regulation. Determining the conservation of protein–protein interactions depends, of course, on how orthologous relationships between proteins are determined. Using a restrictive measure of orthology, such as best bidirectional hits, there are 33 and 45 conserved interactions between the *D. melanogaster* and the Uetz dataset (Uetz et al., 2000) and between *D. melanogaster* and Ito (Ito et al., 2001) dataset, respectively. When dividing by the number of Y2H-interacting *D. melanogaster* proteins whose orthologs are actually present in the *S. cerevisiae*, Uetz and Ito Y2H datasets, there are 24% conserved interactions between *D. melanogaster* and the Uetz dataset and 16% between *D. melanogaster* and the Ito dataset (Table 4.1). These percentages of conserved interactions are substantial when compared with the 26% of interactions that are 'conserved' between the Y2H *S. cerevisiae* datasets. Best bidirectional hits do not necessarily identify functionally equivalent orthologs, especially in the case of gene duplication and varying rates of evolution, and more inclusive measures might identify extra conservation. Using the more inclusive eukaryotic orthologous groups (KOG) [see Clusters of Orthologous Groups (COG) http://www.ncbi.nlm.nih.gov/COG] (Tatusov et al., 2001) to define orthology relationships, the number of 'conserved' interactions between *D. melanogaster* and *S. cerevisiae* increases from 33 to 39 (Uetz dataset) and from 45 to 51 (Ito dataset) (see http://www.cmbi.kun.nl/~huynen/ConservedY2H). Therefore, the issue of how to determine orthology relationships between genomes is becoming less academic because similar functional genomic data for multiple species are available, and we can finally compare the various orthology algorithms for their sensitivity and selectivity.

Vertical genomics – different data from one species

By only analyzing the overlap between genomic data from different species we, of course, ignore biologically relevant, species-specific interactions. To detect such interactions reliably, one can combine different types of genomic data from one species. In doing so one faces several challenges. First, the predictive values of various types of genomic data vary widely not only among different types of genomic data but also within one set of genomic data (e.g. in co-localization data, where the predictive value for protein interaction depends strongly on where in the cell the proteins co-localize (Huh et al., 2003)). Second, datasets tend to

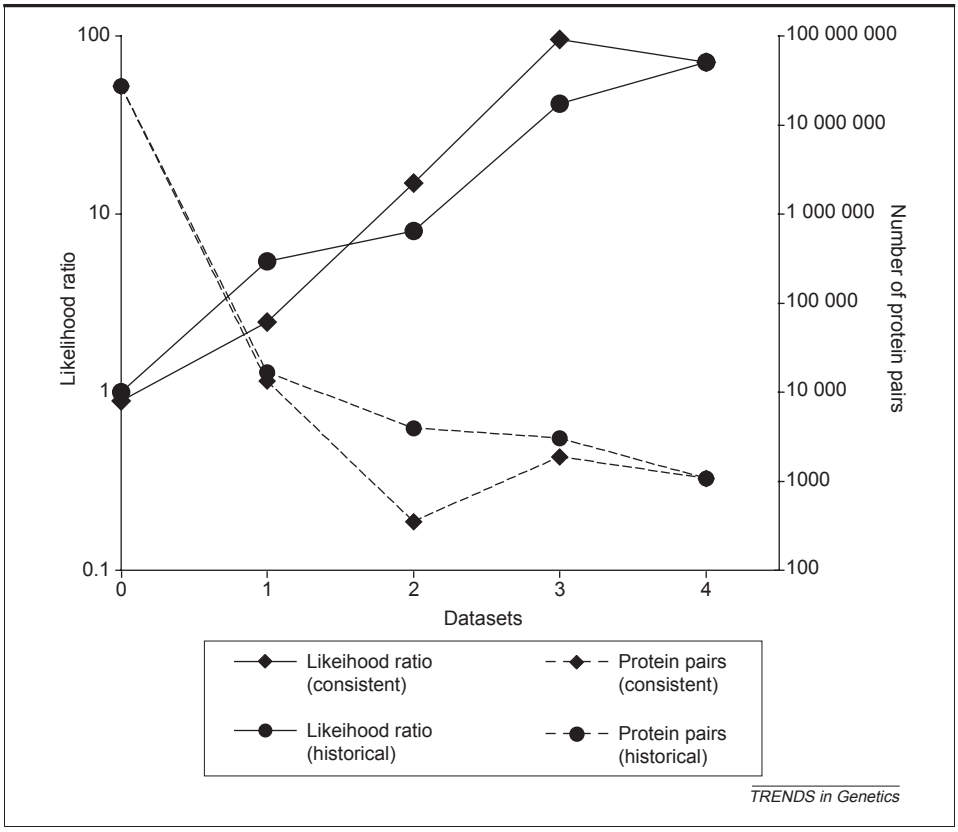


Figure 4.2 | The increase in the likelihood value for the interaction of Plasmodium falciparum proteins observed by consistently correlated expression over four datasets (unbroken lines) and the number of predicted protein pairs at that likelihood value (broken lines). A Bayesian analysis for the two protein and two RNA expression datasets for P. falciparum was performed by calculating pair-wise Pearson-correlation coefficients for all genes in each dataset separately. For the proteomic data, these are based on the number of different peptides detected per protein (Florens et al., 2002) (non-tryptic peptides excluded). Next pairs of genes for each set were divided into two classes: high correlation and low correlation. Finally, for combinations of 'high correlation' and 'low correlation', the relative likelihoods that the proteins interact were estimated based on co-occurrence in Kyoto encyclopedia of genes and genomes (KEGG) maps (http://www.genome.ad.jp/kegg). The 'consistent' line (diamonds) shows the likelihood that proteins interact based on a high correlation in only one dataset (and a low correlation in the other three), a high correlation in two datasets (and a low correlation in the other three) and so on. The line with circles is a historical reconstruction of how the addition of datasets has increased the possibility to predict protein interaction with high likelihood. It is calculated for a high correlation in the first dataset (irrespective of the other three), in the first two (irrespective of the other two) and so on. The order of adding the datasets, from left to right is: proteomics from (Lasonder et al., 2002) and from (Florens et al., 2002), RNA expression from (Le Roch et al., 2003) and from (Bozdech et al., 2003), with correlation thresholds between 'high correlation' and 'low correlation' set at 0.8, 0.75, 0.75 and 0.8, respectively. There are not enough data to divide the correlations into more categories than high and low. For a complete listing of interaction likelihoods of protein pairs, see http://www.cmbi.kun.nl/~huynen/PlasmodiumData. The non-independence of the data is reflected in the saturation of the curve, independent data would, in principle, produce a straight line with the increase in likelihood. The results illustrate the value of having more data for the prediction of protein–protein interactions in P. falciparum, even when those are correlated.

be incomplete: they tend to cover only a fraction of the genes (Yu et al., 2004) (i.e. except for RNA-expression data or genome data). Finally, there are intrinsic correlations between the data, for example, between expression data on the RNA level and on the protein level (Ghaemmaghami et al., 2003).

Recently published approaches tackled these challenges by combining the genomic data in a Bayesian framework (Jansen et al., 2003; Troyanskaya et al., 2003). A Bayesian approach uses a set of known interactions and known non-interactions (e.g. the proteins are in different cell compartments) to estimate how best to combine the various types of data, instead of just 'blindly' combining them as was illustrated previously for the Y2H data. Furthermore, the quality of the predictions is expressed as the likelihood that two proteins interact relative to two randomly chosen proteins and is not an absolute probability as shown in Figure 4.1. When the genomic data are, in principle, independent (e.g. localization and expression data) they are combined in a so-called Naïve Bayesian approach, in which the likelihood that two proteins interact for the separate datasets are multiplied by each other to obtain a combined likelihood. A full Bayesian analysis does not assume the independence of the data and estimates the likelihood by directly comparing combinations of various (binned) values of the data with a set of known interactions and known non-interactions.

A Bayesian network for *Plasmodium falciparum*

We illustrate the Bayesian network approach with an analysis of genomic data for *P. falciparum* for which two gene-expression datasets (Bozdech et al., 2003; Le Roch et al., 2003) and two proteomics datasets (Florens et al., 2002; Lasonder et al., 2002) have been published. Because these data all reflect gene expression, either measured directly as transcript or indirectly as protein, and are therefore not independent, they have to be combined in a full Bayesian framework: the likelihood of protein interaction has to be directly estimated for combinations of correlations between the genes in the separate datasets. Combining the data in this manner leads to an increase in the likelihood of the predictions that can be made in two ways (Figure 4.2): (i) protein interactions that are supported consistently by all datasets are more probable than those that are only supported by one dataset; and (ii) with an increase in data over time the quality of the predictions improves (i.e. having more data enables us to making more likely predictions).

Function prediction in *Plasmodium falciparum*

P. falciparum protein functions can be predicted more reliably using the integrated data. A typical example is PFI0895c, a protein that is homologous to subunit 5 of translation elongation factor 3 (eIF-3 epsilon), which interacts with the ribosome, and is homologous to subunit RPN8 of the 26S proteasome regulatory complex. Both at the RNA and at the protein level, PFI0895c shows a correlated expression with ribosomal proteins L27, L21e and Sa. An annotation of PFI0895c as eIF-3 epsilon appears therefore most likely. Potentially more interesting are predictions for proteins that are specific to the *Plasmodium* genus such as PFI0555c, which is expressed with two proteins that are involved in protein degradation – the aspartic proteinase and drug target (Coombs et al., 2001) PF14_0075 (plasmepsin IV) and the ornithine aminotransferase MAL6P1.91 – suggesting an additional role for PFI0555c in protein degradation.

Outlook

Comparative genomics is a powerful tool to extract reliable predictions from genomic data. To obtain predictions that are amenable to experimental testing, predictions need not only to be reliable but also more specific than 'protein A is involved in process B' or 'protein A interacts with protein C'. One source of information to make predictions more specific is the topology of the predicted interaction networks. Locally densely connected networks reflect stable physical complexes, whereas less connected networks correlate with signaling pathways and transient interactions (Pereira-Leal et al., 2004; Spirin and Mirny, 2003). With the avalanche of genomic data, instead of merely determining the overlaps, scientists will be in a position to determine the differences and extract biological meaning from those differences. In contrast to stable complexes, transiently interacting proteins appear not to show correlated expression (Jansen et al., 2002). Alternatively, one might be able to find metabolic pathways in consistently co-expressed but not physically interacting proteins. We will undoubtedly see that more creative combinations of genomic data will increase the specificity of genomics-based protein-function prediction, although whether specific experimental testing of protein function prediction will ever catch up with the large number of function predictions remains to be seen.

Chapter 5

Combinatorial gene regulation in *Plasmodium falciparum*

Vera van Noort and Martijn A. Huynen

Trends in Genetics 22, 73-78 (2006)

Combinatorial gene regulation in *Plasmodium falciparum*

Abstract

The malaria parasite *Plasmodium falciparum* has a complicated life cycle with large variations in its gene expression pattern, but it contains relatively few specific transcriptional regulators. To elucidate this paradox, we identified regulatory sequences, using an approach that integrates the sequence conservation among species and the correlation in mRNA expression within a species. Our analysis identified several DNA sequence motifs that are associated with mRNA expression, two of which were previously determined experimentally. We found more putative regulatory sequences per gene in *P. falciparum* than in other eukaryotes, such as yeast. We propose that *Plasmodium* uses the few regulatory proteins it has in a combinatorial approach for gene regulation, explaining the relative paucity in regulatory proteins.

Introduction

Half of the population of the world lives in areas where malaria is endemic, causing the death of up to three million people annually. The most lethal form of human malaria is caused by infection with the protozoan parasite *Plasmodium falciparum*. The genome sequence of this eukaryotic organism in addition to mRNA and protein expression data are publicly available; however, the gene-regulatory processes governing the development of the parasite are poorly understood. Proteomics (Florens et al., 2002; Lasonder et al., 2002) and mRNA expression data (Bozdech et al., 2003; Le Roch et al., 2003) show that *P. falciparum* has major variations in gene expression levels throughout its life cycle. Furthermore, transcription levels are influenced by environmental factors such as temperature and glucose concentration(Fang and McCutchan, 2002; Fang et al., 2004). However, the *Plasmodium* genome seems to encode relatively few proteins that are homologous to transcription factors found in other eukaryotes; these transcription factors are expected to contribute to gene-specific transcriptional regulation (Coulson et al., 2004). How the parasite manages to control the timing of gene expression correctly taking into account the requirements of the cell remains elusive. It has been proposed that histone-modifications or post-transcriptional mechanisms have a larger effect on gene expression than transcriptional regulation in *Plasmodium*. Recently, six genes were shown to contain sequences that might be implicated in translational repression (Hall et al., 2005).

The target sequences of transcription regulators in *P. falciparum* are largely unknown. Two methods exist to detect cis-regulatory elements by bioinformatics approaches. The first method determines shared sequence motifs in upstream regions of genes that have

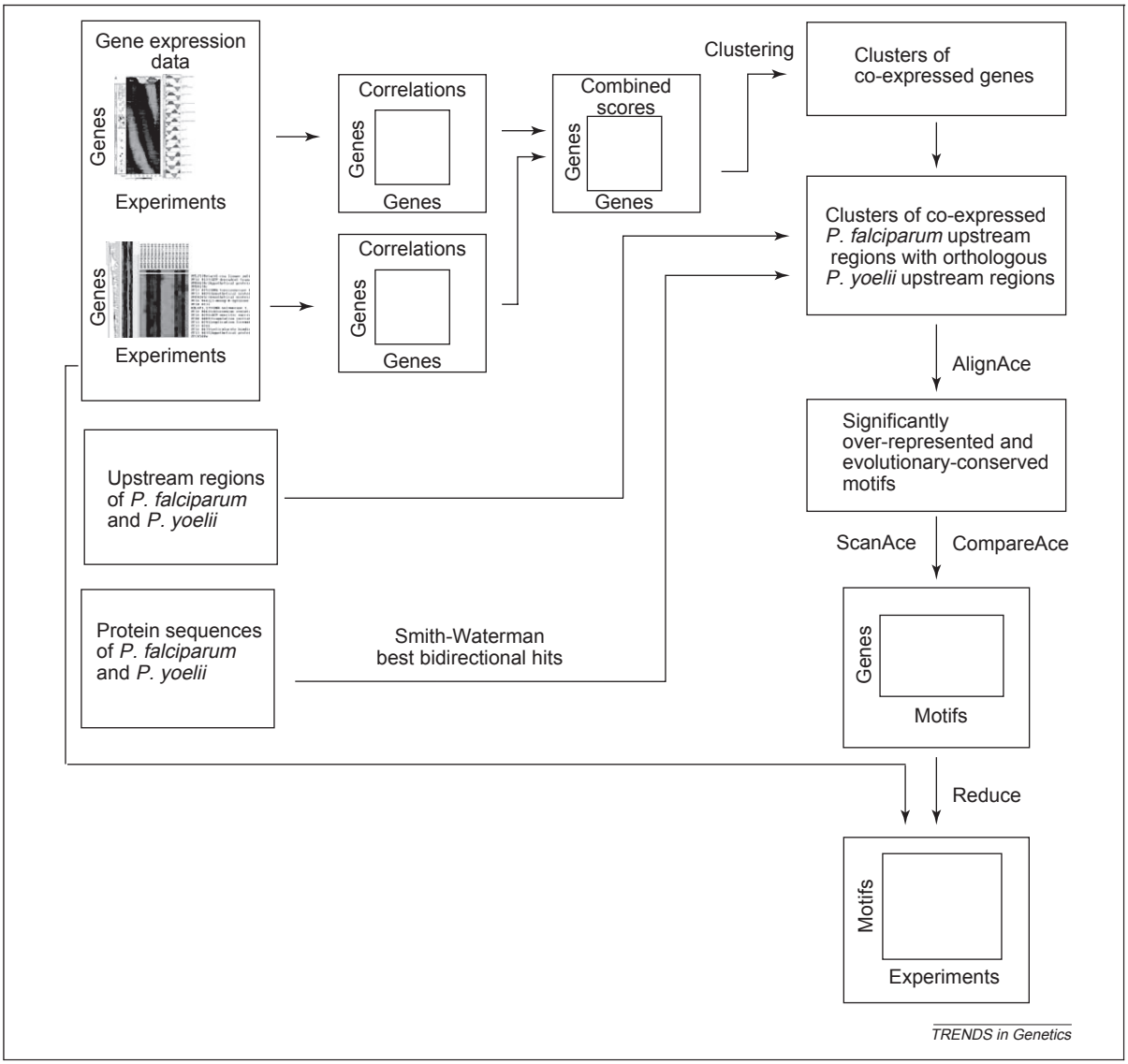


Figure 5.1 | Cis-regulatory motif detection. First, the correlations between gene pairs were calculated on the basis of two mRNA expression data sets (Bozdech et al., 2003; Le Roch et al., 2003). Genes were then clustered, and the clusters that contained at least 20 genes were considered for further analysis. The squares indicate the calculated data (correlation and scoring matrices). Next, the regions that were 1-kb upstream from the co-expressed Plasmodium falciparum genes and their orthologous genes in Plasmodium yoelii were selected. The clusters of upstream regions were subsequently used as input for the AlignAce program (Hughes et al., 2000; Roth et al., 1998), which finds over-represented motifs by a Gibbs-sampling algorithm. The upstream regions of all P. falciparum genes were scanned for the presence of over-represented motifs, resulting in a scoring matrix of 5334 genes by 79 motifs. To obtain motifs that correlated with the expression data, we used the multivariate regression approach with forward motif selection used in Reduce (Bussemaker et al., 2001). We included motifs until $P<0.01$ of the most significant motif. Time courses (T-values) were calculated for all significant motifs.

similar expression patterns or similar functions (Roth et al., 1998; van Helden et al., 1998). These motifs have, in several cases, been shown to be target sites. The second method is 'phylogenetic footprinting', in which conserved sequences among multiple species in non-coding DNA can indicate regulatory sites (Cliften et al., 2003). Cis-regulatory elements are conserved at a significantly greater level than non-functional DNA among genomes that are as distant as human and mouse genomes (Liu et al., 2004). The evolutionary divergance between the rodent parasite *Plasmodium y. yoelii* and the human parasite *P. falciparum* is approximately the same as that between human and mouse (Carlton et al., 2002), leading to the expectation that cis-regulatory elements will also be conserved between these two *Plasmodium* species. In a preliminary study, the AlignAce program was used to find cis-regulatory elements in *P. falciparum* in upstream regions of genes encoding heat shock proteins (Militello et al., 2004), leading to the identification of the G-box element. Comparison among different *Plasmodium* species revealed that this element is conserved.

The extreme AT-richness of *Plasmodium* intergenic regions makes it difficult to identify putative regulatory elements by either phylogenetic footprinting or over-representation in functionally related genes. Therefore, we integrated the two approaches to identify these elements (i.e. we used both clusters of co-expressed genes in *P. falciparum* and the evolutionary sequence conservation between *P. y. yoelii* and *P. falciparum*). We found 12 putative regulatory motifs. Based on our results, we hypothesize that *P. falciparum* uses a greater number of transcriptional regulatory sites per gene in a combinatorial fashion compared with other eukaryotic species, such as *Saccharomyces cerevisiae*.

Integrating evolutionary conservation with expression correlation

Our method for putative regulatory-element detection looks simultaneously for motifs that correlate with mRNA expression profiles and have evolutionary sequence conservation. An overview of the method is given in Figure 5.1, details can be found in the Supplementary Material. First, we calculated similarity in expression of gene pairs based on two expression data sets (Bozdech et al., 2003; Le Roch et al., 2003) by multiplying the gene–gene correlations from the individual data sets. Then we clustered *Plasmodium* genes based on the combined co-expression scores. Finally, these *P. falciparum* gene clusters were combined with their *P. y. yoelii* orthologs to find conserved motifs in the upstream regions using the AlignAce program. Note that we did not make alignments of upstream regions but instead used the Gibbs sampler from AlignAce (Roth et al., 1998) to find motifs that correlate simultaneously with expression and with evolutionary conservation.

Correlation of motifs with expression data

Using the different co-expression clusters, we found 79 over-represented motifs. The upstream regions of all *P. falciparum* genes were scanned for presence of the motifs. For each motif cluster, we took the greatest score for each gene resulting in a scoring matrix of 5334 genes by 79 motifs. Among these motifs we expected that some are functional and, therefore, correlate with specific expression patterns, whereas others might just be over-represented in the whole genome or occur because of other biases, specifically in the AT-rich genome of *P. falciparum*. Therefore, we used a second algorithm, Reduce, to calculate the correlation (r) between the motif scores and expression levels (Bussemaker et al., 2001), to

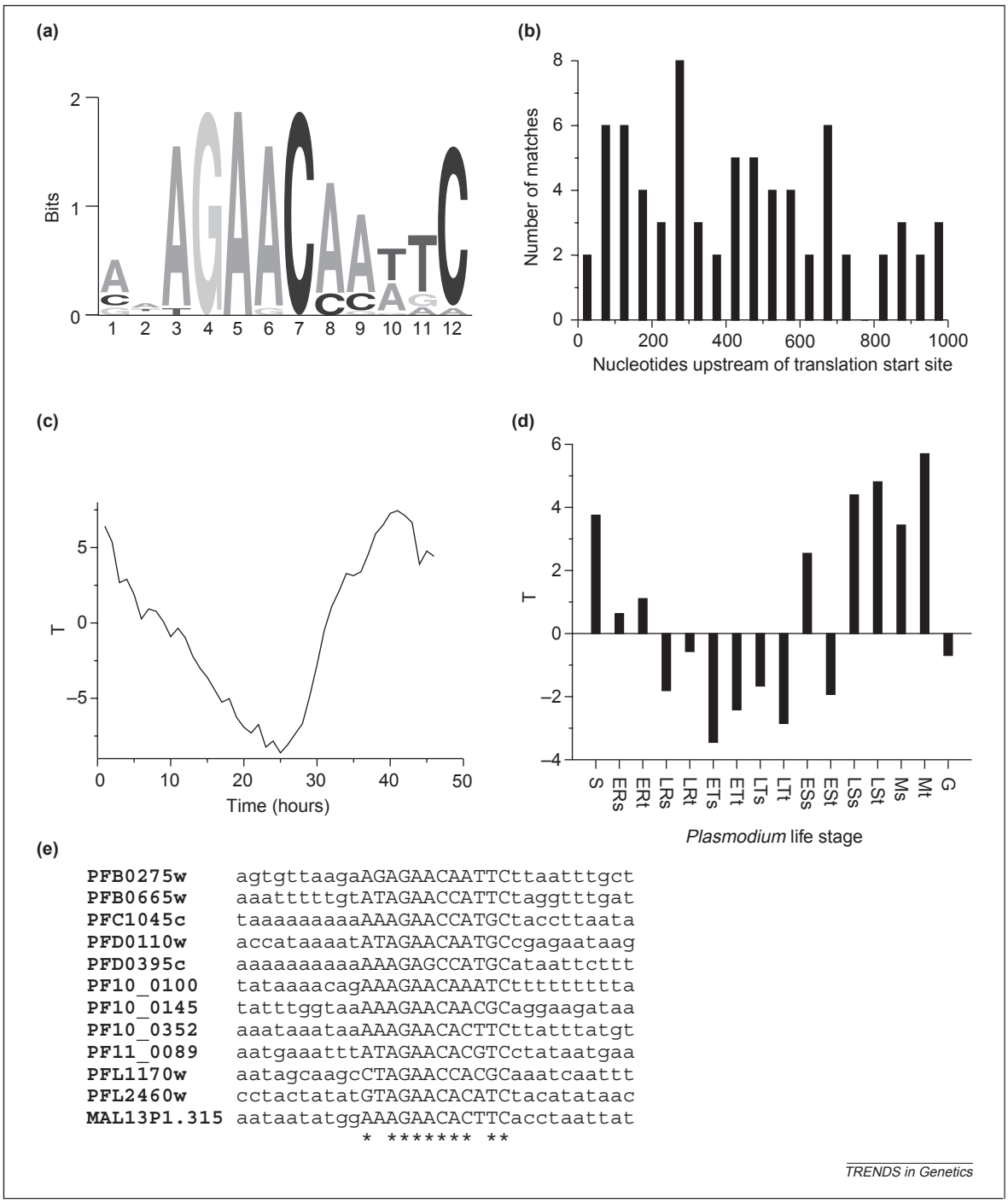


Figure 5.2 | The late-schizont motif. (a) Logo of the late-schizont regulatory motif. The height per position represents the information content and the height of the letters indicates the frequency. (b) Positions of non-overlapping matches counted in 50-nt bins. Most matches of the late-schizont motif are 250–300 nt upstream of the translation start site (position preference $P<0.05$). (c) Time course (T) for the late-schizont motif in the Bozdech data set. (d) Time course for the late-schizont motif in the LeRoch data set. (e) The alignment of upstream regions of genes expressed in late schizonts that contain the motif. The asterisks indicate conserved positions. The region matching the motif is in capital letters. Abbreviations: ER, early ring; ES, early schizont; ET, early trophozoite; G, Gametocyte; LR, late ring; LS, late schizont; LT, late trophozoite; M, Merozoite; S, sporozoite; s, sorbitol-synchronized parasite; t, temperature-synchronized parasite.

obtain potentially functional motifs. This method ensures that if similar motifs correlate with the same gene expression levels, we will only identify the motif that has the strongest correlation with these gene expression levels. We found 12 putative regulatory motifs [i.e. motifs that significantly correlated ($P<0.01$) with mRNA expression]. Note that out of 79 over-represented motifs, only 12 correlated with mRNA expression. Indeed by just detecting over-representation, we identified motifs that were the result of biases in the genome.

Previously determined sequence motifs

The sequence logos of each of these motifs and correlations with expression data can be found in Table 1 of the Supplementary Material. There is little experimental data available about expression-related sequence motifs in *P. falciparum*. Two of the motifs we identified had been experimentally determined previously. Motif 5 contains consensus sequence TGTATATATG and is correlated with upregulation in schizonts in mRNA expression data sets and in gametocytes in the LeRoch data set. Motif 5 is similar to and present in the same genes as the sequence TGTAT(G/A)TG, which was found to regulate var genes in an experimental study (Calderwood et al., 2003). We also observed a poly(dAdT) repeat (motif 11), which correlated with upregulation in gametocytes, this motif was recently found to regulate calmodulin activity (Polson and Blackman, 2005).

Two known regulatory sequences were not present in our results. The first is an experimentally determined sequence associated with sexual and early mosquito stages: the recognition site for the PAF-1 transcription factor (Dechering et al., 1999). The second motif that we did not retrieve is the CCAAT box, although this is to be expected because the *Plasmodium* genome encodes the complete CCAAT-box-binding complex (Coulson et al., 2004). Additional mRNA expression data will be needed for the identification of these sequence motifs.

Newly discovered motifs

In addition to motif 5, we found two more T(G/A) repeat motifs (motifs 3 and 4) that also correlate with mRNA expression in schizonts and gametocytes. We found other important motifs including oligo(dA)oligo(dT) repeat (motif 1), correlating with gene upregulation in the ring stage and a poly(dG)poly(dA) motif (motif 2), correlating with gene upregulation in the trophozoite stage. Figure 5.2 shows the late-schizont motif, a newly discovered, putative regulatory motif that correlates with upregulation of 72 genes in late schizonts (Figure 5.2c,d). The late-schizont motif occurs preferentially ($P<0.05$) between 250 and 300 nucleotides upstream of the translation start site (Figure 5.2b). Binding of transcriptional regulators in yeast occurs predominantly at ~180 bases upstream of the start codon (Harbison et al., 2004). Thus, we expect that the late-schizont motif also constitutes a transcription-factor-binding site. Similar to the late-schizont motif, all motifs have a significant position preference relative to the translation start site. The same motifs that are correlated with expression in the Bozdech data set also correlate with expression in the LeRoch data set. Both data sets describe asexual intra-erythrocytic blood stages. The LeRoch data set describes three additional parasite life stages, enabling the discovery of additional motifs.

Functional significance of discovered motifs

In the absence of large amounts of experimental data on transcription-factor-binding sites in P. falciparum, we used other data and randomizations to asses the functional significance

of the discovered motifs.

First, we examined the robustness of our results by applying other strategies to find cis-regulatory motifs. Rather than examining co-expressed genes, we analyzed upstream regions of genes with similar functions (occurring in one pathway as defined by the Kyoto encyclopedia of genes and genomes (KEGG) data base (Ogata et al., 1999)), we combined them with the upstream regions of their orthologs in *P. y. yoelii* and used those as input sets for AlignAce to detect over-represented motifs. These motifs are clustered and correlated with expression patterns. Logos (Crooks et al., 2004) of motifs with the greatest correlations with expression are shown in Table 2 of the Supplementary Material. We found similar motifs with this approach, confirming the functional relevance of the discovered motifs. A few forms of the T(G/A) repeat (motif 15, 17, 18 and 27), the two different AT-rich repeats [poly(dAdT), motif 28; oligo(dA)oligo(dT) repeat, motif 13] and the poly(dG)poly(dA) motif (motif 14) seem to regulate a functionally coherent set of genes. However, the variation that can be explained by these motifs is less than that explained by motifs in clusters of co-expressed genes (Figure 5.3a).

Second, we found that simpler methods do not give better or even as good results as our approach. We performed an exhaustive search for oligomers, up to 7 nt in length, in the upstream regions of P. falciparum and examined their predictive value. These oligomers gave a lower correlation with the expression data than the over-represented motifs; for example, the top scoring oligomer, GACCGC, only has a maximum r^2 value of 0.0125, whereas our top scoring motif (motif 3) has a score of 0.108.

Third, to examine the value of including the upstream regions from the second species in the study, we repeated the analysis without *P. y. yoelii*, by first determining over-represented motifs in upstream regions of *P. falciparum* and then examining their correlation with gene expression. This resulted in motifs that were not correlated with the expression data (Figure 5.3a), underscoring the value of combining sequence conservation with co-expression data in determining regulatory elements.

Finally, we verified that the statistical significance of the correlations of motifs with mRNA expression corresponds to a real signal in the upstream regions of genes and not to other biases in intergenic regions of *P. falciparum* by randomizing our data. All expression profiles were randomly reassigned to the genes, effectively detaching upstream regions of genes from the gene expression data. Next, we re-clustered the genes according to their new expression profiles and repeated the motif discovery procedure. This resulted in sequence motifs that had little correlation with the (still randomized) expression profiles (Figure 5.3a), indicating that in the *Plasmodium* genome there are DNA sequences in the upstream regions of genes that correlate with mRNA expression. Whether these DNA sequences are transcription-factor-binding sites will have to be solved experimentally. It is possible that the sequence elements are related to mRNA stability or chromosome accessibility.

Combinatorial gene regulation

To elucidate the paradox of the small number of regulators encoded in the *Plasmodium* genome, we counted the number of regulatory elements (motif clusters that correlate significantly with expression) per gene for yeast and *Plasmodium* (Figure 5.3b). Most *Plasmodium* genes have four or five different regulatory elements in their upstream region. This contrasts

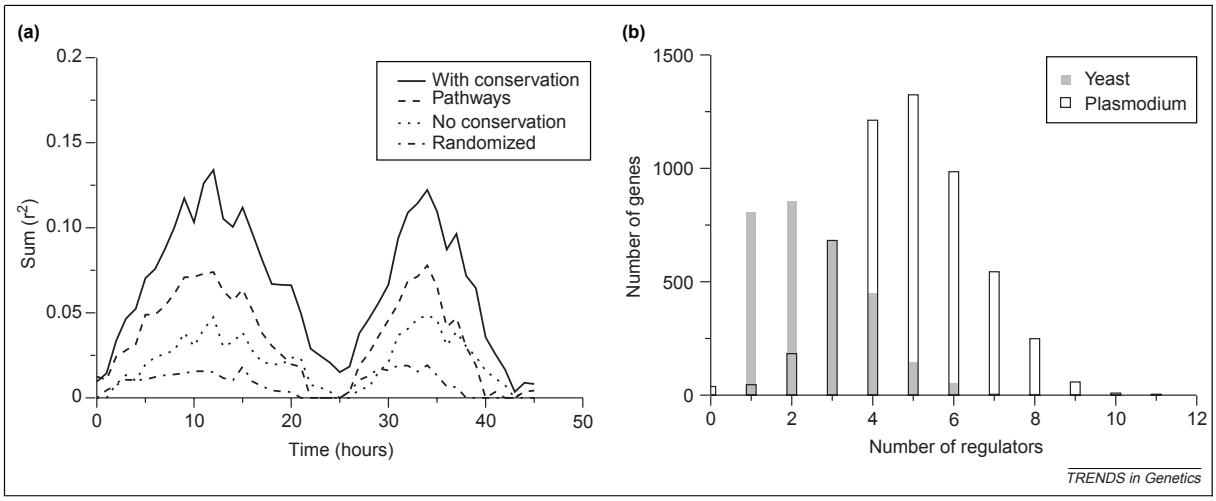


Figure 5.3 | Regulatory motifs in *Plasmodium falciparum*. **(a) The variation in gene expression that can be explained by motifs. For each experiment (shown on the x-axis), we calculated the total amount of variation (r2) in expression levels in the Bozdech data set that can be explained by motifs identified using four different methods [motifs found with the procedure as depicted in Figure 1 (unbroken line); motifs found without conservation (dotted line); motifs found using KEGG pathways (dashed line); motifs found with upstream regions and expression profiles randomized relative to each other (dot-dashed line)]. The greatest level of variation that can be explained by motif scores was obtained by the regulatory-element-detection method that integrates both mRNA expression and evolutionary sequence conservation. (b) The number of regulators per gene. In yeast, the number of unique oligomers correlating with expression was obtained from (Bussemaker et al., 2001) and counted in 1000-bp regions upstream of yeast genes that do not overlap with coding regions (grey). For *Plasmodium*, the number of different motifs that correlated significantly with expression was counted in 1000-bp regions upstream of *Plasmodium falciparum* overlapping genes (white).**

strongly with the situation in yeast, where most genes are regulated by only one or two regulators (Figure 5.3b). Chromatin immunoprecipitation (ChIP-on-chip) data from *Plasmodium* are not yet available, therefore, we chose to compare the distribution of regulators per gene with the results of a similar computational method in yeast (Figure 5.3b). ChIP-on-chip data also show that the vast majority of yeast genes have only one regulator binding to their promoter region (Lee et al., 2002). However, in *Plasmodium* it seems that fewer regulators are used in different combinations of five elements per promoter to obtain the same level of diversity in expression profiles. A simple calculation shows that with ten regulatory proteins and five elements per gene, (10 choose 5) = 252 different combinations can be made. If promoter sequences contain only one regulatory element, then 252 instead of ten regulatory proteins would be needed to obtain the same number of expression profiles. If they contain two elements per gene, then 23 regulatory proteins would be needed.

The most abundant combination of regulatory elements is 1+4+8+10+11, a combination of elements some of which have opposite effects on gene expression. For example, motifs 1 and 4 have opposite effects in all experiments (Table 1 in the Supplementary Material) and 775 genes have a combination of these motifs. This leads to the hypothesis that *Plasmodium* uses combinatorial effects of gene regulators, exploiting the possibilities of the relatively few regulators that it possesses (Coulson et al., 2004). Instead of using one regulatory protein for each expression profile, different combinations of regulators are employed to obtain a variety of expression profiles. Experimental results support this hypothesis. First, studies of the promoter of GBP130 showed that this gene was most likely to be regulated by multiple, possibly different nuclear factors (Horrocks and Lanzer, 1999). Furthermore, a study of var genes showed that silencing occurs through the cooperative action of multiple

sequence elements (Deitsch et al., 2001). Finally, a study of the Polδ promoter revealed regions that have both positive and negative effects on gene expression (Porter, 2002).

Concluding remarks

We have identified DNA motifs in the upstream regions of *Plasmodium falciparum* genes that significantly correlate with mRNA expression. To find these motifs it was necessary to integrate phylogenetic footprinting techniques with the over-representation of motifs in co-expressed genes. The results provide an explanation for the paradox between the large variation in *P. falciparum* gene expression and the reported paucity of specific transcription factors. It can be explained by a combinatorial mode of gene regulation, in which every gene is regulated by multiple factors. Our results suggest that this is the general mode of transcriptional regulation in *Plasmodium*; that is combinations of regulatory motifs contribute to overall promoter activity.

Materials and Methods

Co-expression clusters

For two mRNA expression datasets (Bozdech et al., 2003; Le Roch et al., 2003) correlation coefficients between all gene-pairs were calculated based on the log ratio expression values. The correlations per expression dataset were renormalized to coefficients between 0 and 1. Subsequently the two coefficients per gene pair arriving from two datasets were multiplied with each other in order to obtain a single expression similarity score for each gene pair. Average linkage hierarchical clustering was applied and gene clusters were identified using a cut-off. We chose the cut-off such that we obtained 20 clusters of at least 20 genes. Only clusters of at least 20 genes were considered for further analysis.

Orthologs

Protein sequences of *P. falciparum* and *P. y. yoelii* were downloaded from PlasmoDB(Bahl et al., 2003). Smith-Waterman searches(Bahl et al., 2003) of all *P. falciparum* proteins were done against all *P. y. yoelii* proteins. The hits with the lowest E-value were determined "Best Hits". Orthologs were assigned by taking Bidirectional Best Hits between the two protein sets. In case of two Best Hits with the same E-value, two proteins from one species were assigned orthologous to one protein from another species. In this way, orthologous relations were assigned between 3998 *P. falciparum* proteins and 4036 *P. y. yoelii* proteins.

Motif detection

The genomes and genome annotations of *P. falciparum* and *P. y. yoelii* were obtained from PlasmoDB(Bahl et al., 2003). For the genes in the 20 co-expressed clusters, upstream regions of 1kb were selected. In case of overlap with annotated protein coding genes, only the non-coding part was taken. In the same manner upstream regions of orthologs in *P. y. yoelii* were obtained. Upstream regions of co-expressed genes together with upstream regions of their orthologs were used as input for the AlignAce program(Hughes et al., 2000; Roth et al., 1998), that finds overrepresented motifs by a Gibbs sampling algorithm. The default parameters were used, except for the GC content which was set to 0.2; the average GC content of

P. falciparum and *P. y. yoelii* noncoding DNA. The upstream regions of all *P. falciparum* genes were scanned for presence of the resulting motifs using ScanAce with GC content set to 0.13 (the GC content of *P. falciparum*) and standard deviation of 1. The maximum number of returned sites was set such that all matches of motifs in upstream regions were returned. ScanAce returns alignment scores with all motifs for each match. Using the different co-expression clusters, we found 187 overrepresented motifs. Similarity between these motifs was calculated using CompareAce, that returns the highest Pearson correlation (c) between pairs of motifs. Similar motifs (c > 0.7) were clustered together resulting in a total of 79 motif clusters. We obtained one score for each gene by taking the highest ScanAce score with a particular motif cluster for that gene, resulting in a scoring matrix of 5334 genes by 79 motifs.

Motif significance

To obtain motifs with independent correlations with the expression data, we used the multivariate regression approach with forward motif selection used in REDUCE(Bussemaker et al., 2001), using the set of matrices discovered by our conservation-based procedure as input rather than all oligonucleotides up to a given length and replacing motif counts by motif scores (See Motif detection). The correlation between motif score and expression level was calculated for all expression data points in both datasets. The correlation of the most significant motif was subtracted from the expression data, after which correlations between motifs and expression data were recalculated. The correlation is transformed into a T-value $T = r\sqrt{G}$ where G is the number of genes and r the Pearson correlation. P-values were calculated using the Bonferroni correction to compensate for multiple testing. We included motifs until P > 0.01 of the most significant motif.

Time-courses of motifs

Time courses (T-values) were calculated for all significant motifs for all datapoints. If the motif scoring data would be presence/absence this would result in the average profile of all genes bearing the motif, but in this case profiles of genes that score higher on the motif get a higher weight in calculating the average profile.

Position preference

The probability of observing m or more sites out of a possible t in a 50 bp window of a 1000 bp region is determined by the formula:

$$P = \sum_{i=m}^{t} \binom{t}{i} \left(\frac{w}{s}\right)^{i} \left(1 - \frac{w}{s}\right)^{t-i}$$

where w = 50 and s = 1000 (Hughes et al., 2000).

Conservation score

A conservation score is calculated by dividing the number of *P. falciparum* genes in a co-expression cluster that contain a specific motif of wich the ortholog in *P. yoelii* also contains the motif by the total number of *P. falciparum* genes in a co-expression cluster that contain a specific motif and have an ortholog in *P. yoelii*. Conservation scores are given in Table 1 and

Table 2. In parentheses is the total number of *P. falciparum* genes in a co-expression cluster that contain a specific motif and have an ortholog in *P. yoelii*. For randomized orthology relations 86% of the conservation scores are 0, the maximum random conservation score is 0.16.

Total variation explained by motifs

For each experiment (x-axis) the total amount of variation (r^2) in expression levels in te Bozdech dataset that can be explained by motif scores is given for motifs found using upstream regions of both *P. yoelii* and *P. falciparum* in co-expression clusters; with the same procedure but without taking the upstream regions of *P. y. yoelii* into account; for motifs found using upstream regions of falciparum and yoelii genes that are together in KEGG pathways; with the procedure including sequence conservation but with upstream regions and expression profiles randomized relative to each other.

Number of regulators per gene

For yeast, the number of unique oligomers correlating with expression were taken from ref (Bussemaker et al., 2001) and counted in 1000 bp upstream regions of yeast genes not overlapping with coding regions. Genes with upstream regions that overlapped completely with coding regions were not considered. For *Plasmodium*, the number of different motifs significantly correlating with expression (12 motifs from Table 1) were counted in 1000 bp upstream regions of *Plasmodium falciparum* genes not overlapping with coding regions.

Acknowledgements

We thank Harmen Bussemaker for careful reading of the article and Chris Ponting for useful discussions. This work was supported in part by a grant from the Netherlands Organization for Scientific Research (NWO).

Chapter 6

Exploration of the omics evidence landscape to distinguish metabolic from physical interactions

Vera van Noort, Berend Snel and Martijn A. Huynen

Submitted

Exploration of the omics evidence landscape to distinguish metabolic from physical interactions

Abstract

In the post-genomic era various functional genomics, proteomics and computational techniques have been developed to elucidate the protein interaction network. While some of these techniques are specific for a certain type of interaction, most predict a mixture of interactions. Currently no method exists that can specifically find proteins that function in the same metabolic pathway without also retrieving proteins that are part of the same protein complex. We here fill this gap by constructing an "omics evidence landscape" that combines all sources of evidence for protein interactions from various types of omics data. We explore this evidence landscape to identify areas with only metabolic or only physical interactions, allowing us to specifically predict the nature of new interactions in these areas. The information from the evidence landscape, allows us to survey a protein interaction network of *Saccharomyces cerevisiae* with qualitative information about the interactions.

Introduction

Genome sequencing projects have resulted in the listing of all the protein coding and RNA genes for a large number of organisms. To determine which of these genes function together, interaction networks have been elucidated using a plethora of omics (genome-scale) techniques. However, for the biological interpretation of such networks and the prioritization of experimental verification, we do not only need to know whether proteins interact, but also how they interact.

The most straightforward type of interaction is a physical interaction in which two proteins actually bind to each other. Physical protein interactions have been studied by genetic, biochemical and biophysical techniques, but also by high-throughput interaction-detection methods. Enormous amounts of data have been collected by yeast-2-hybrid assays (Ito et al., 2001; Uetz et al., 2000) and complex purification methods (Gavin et al., 2006; Gavin et al., 2002; Ho et al., 2002; Krogan et al., 2006). More qualitative information on the nature of physical interactions of so-called hub proteins was collected by overlapping the interactions with co-expression data (Han et al., 2004).

Another type of interaction exists between proteins that are part of the same pathway. In this interaction the proteins do not directly bind to each other but, for example, pass metabolites or information to each other. Metabolic interactions, in which proteins are part of the same metabolic pathway, are the clearest exponent of these pathway interactions. No method exists that exclusively detects metabolic interactions, even though they are detected by certain methods together with other interactions. Correlated messenger RNA expression

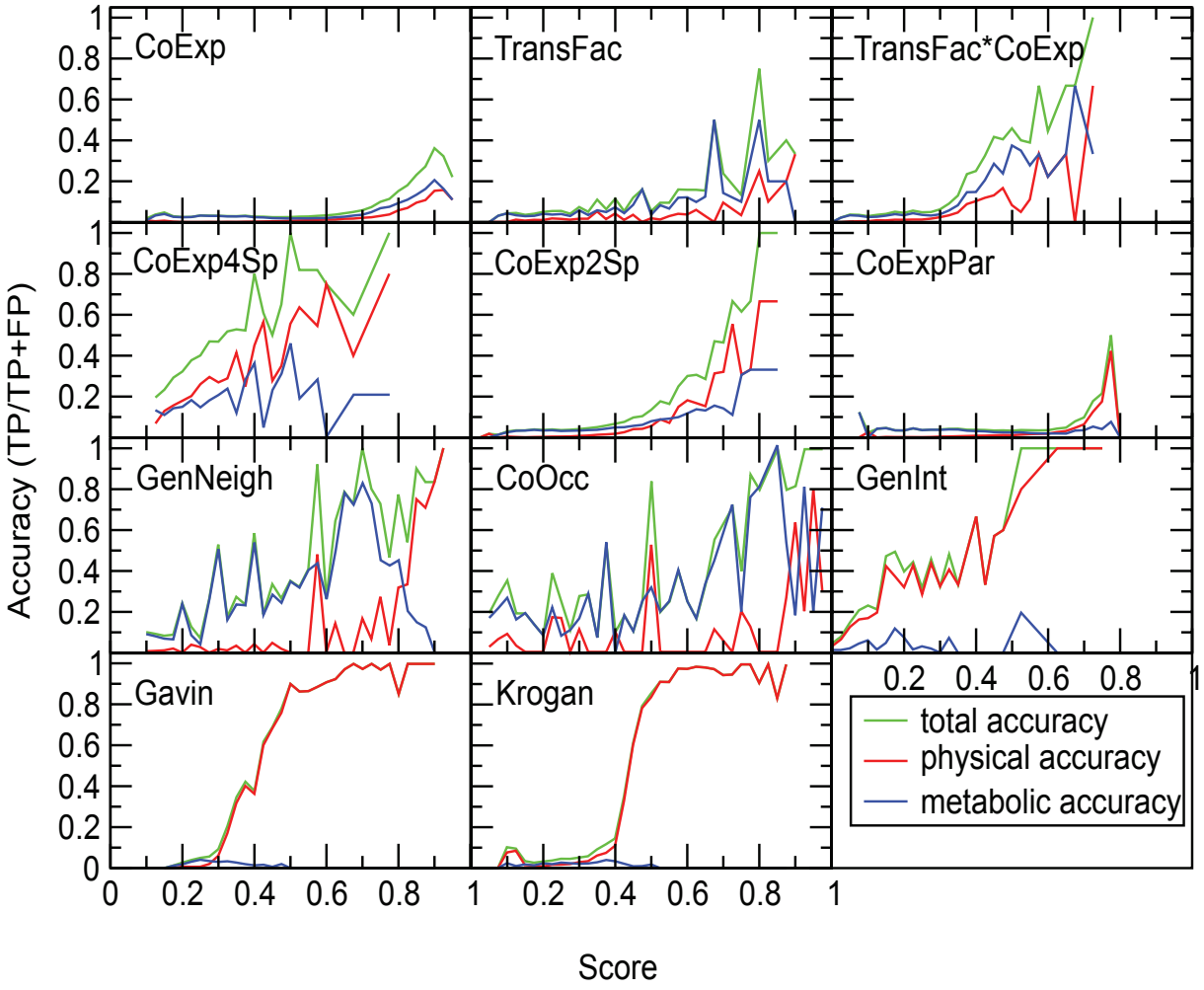


Figure 6.1 | Score-accuracy plots of individual datasets. On the x-axis is the score for that data set, on the y-axis the accuracy. Data were binned, bins were chosen such that each bin contains at least five gene pairs. The green lines indicate the total accuracy, meaning the total number of True Positives divided by the total number of True Positives plus False Positives in that bin. The red lines indicate the accuracy on the protein complex reference set, being the number of True Positives in the complex reference set divided by the number of True Positives and False positives in both reference sets. The blue lines indicate the accuracy on the metabolic reference set, being the number of True Positives in the metabolic reference set divided by the number of True Positives and False positives in both reference sets. **a** Correlated mRNA expression (CoExp) **b** Shared binding of transcription factors (TransFac) **c** Co-regulation (TransFac*CoExp) **d** Conserved co-expression between four species (CoExp4Sp) **e** conserved co-expression between two species (CoExp2Sp) **f** Paralogous conserved co-expression (CoExpPar) **g** Gene neighborhood conservation (GenNeigh) **h** Correlated phylogenetic profiles (CoOcc) **i** Shared genetic interactions (GenInt) **j** Protein-protein interactions (Gavin) **k** Protein-protein interactions (Krogan) For **j** and **k** the protein-pairs that are never co-purified and thus have a socio-affinity score of 0 are in bin 0.2.

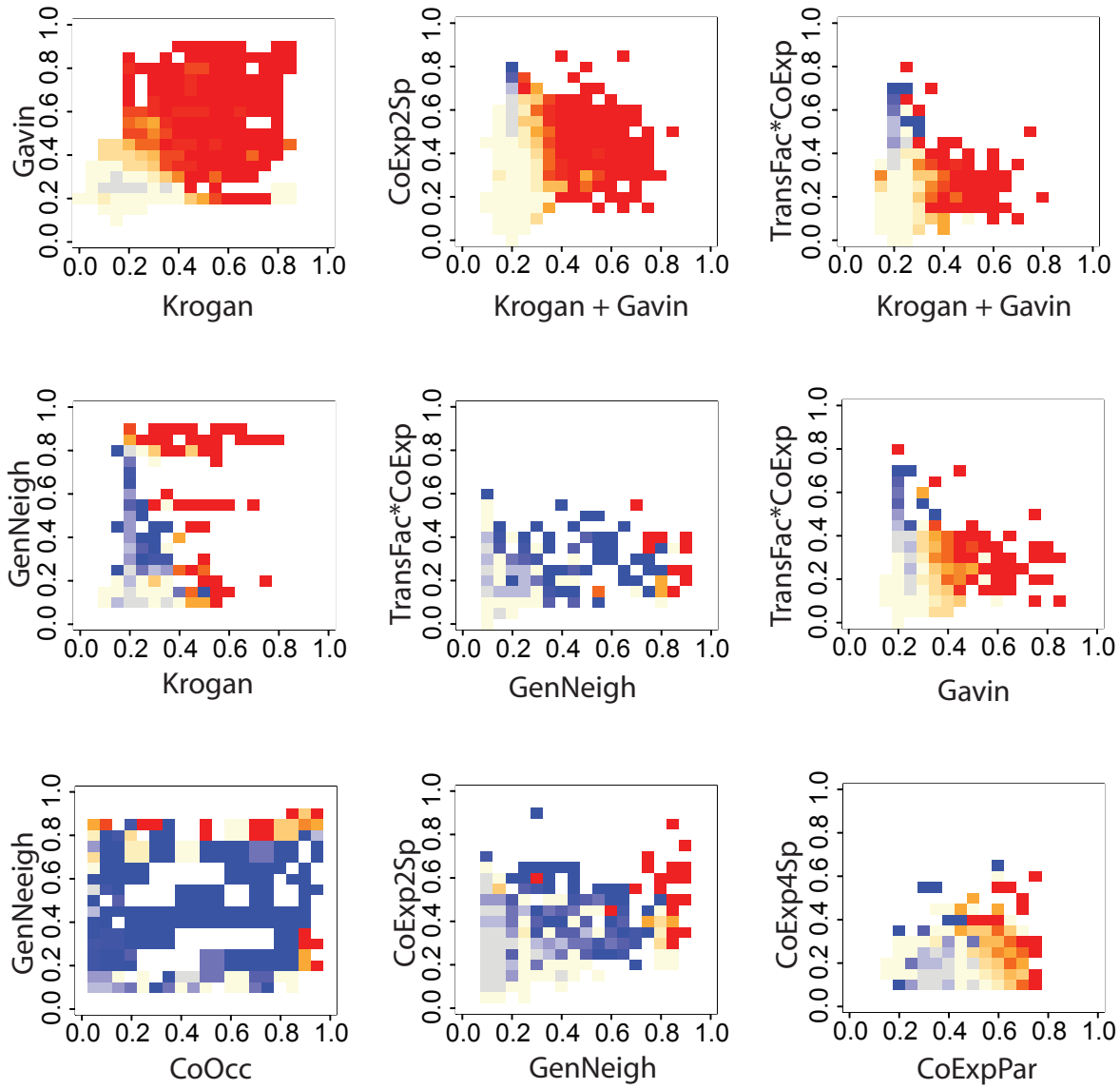


Figure 6.2 | Differential accuracy in the evidence landscape. In each panel the x-axis indicates the score in the first dataset, the y-axis the score in the second set. The color scheme is based on Differential Accuracy, being the accuracy on the metabolic reference set minus the accuracy on the protein complex reference set. Differential Accuracy 1 is dark blue, 0 is yellow and -1 is red, parts that contain no gene pairs are white. The blue parts of the landscapes are regions were there are only metabolic interactions, whereas in the red parts there are only physical interactions. a Protein-protein interactions (Krogan) versus Protein-protein interactions (Gavin). b Protein-protein interactions (sum Krogan Gavin) versus conserved co-expression (CoExp2Sp). c Protein-protein interactions (sum Krogan Gavin) versus co-regulation (TransFac*CoExp). d Protein-protein interactions (Krogan) versus Gene neighborhood conservation (GenNeigh). e Gene neighborhood conservation (GenNeigh) versus Co-regulation (TransFac*CoExp). f Protein-protein interactions (Gavin) versus Co-regulation (TransFac*CoExp). g Correlated phylogenetic profiles (CoOcc) versus gene neighborhood conservation (GenNeigh). h Gene neighborhood conservation (GenNeigh) versus conserved co-expression (CoExp2Sp). i Paralogous conserved co-expression (CoExpPar) versus conserved co-expression (CoExp4Sp).

profiles and genetic interaction data (Tong et al., 2004), as well as gene fusion, conserved gene neighborhood and gene co-occurrences or phylogenetic profiles are indicative of both physical as well as metabolic interactions (Huynen et al., 2000; Jensen et al., 2004; Kelley and Ideker, 2005). Metabolic interactions can also be predicted together with physical interactions from integrated co-expression data between species (Stuart et al., 2003; van Noort et al., 2003) or from the integration of co-expression with the sharing of transcription factors (Snel et al., 2004).

The plethora of omics data sets that are available have been successfully integrated in order to reduce experimental noise and obtain reliable predictions about protein-protein interactions (Beyer et al., 2006; Huynen et al., 2004; Jansen et al., 2003; Joyce and Palsson, 2006; Troyanskaya et al., 2003; von Mering et al., 2003). The retrieval of qualitative information on the nature of the interactions has received little focus, despite the usefulness of such qualitative information for the biological interpretation of interactions. We here present an integration that does differentiate one type of interaction from another, allowing such qualitative interpretation of the predicted interactions between proteins.

We choose to integrate omics data sets for the budding yeast *S. cerevisiae* because of the availability of both many high quality genomics data as well as classical knowledge about its protein functions. To be able to distinguish physical interactions from metabolic ones we construct two separate reference sets: one for physical interactions and one for metabolic interactions. We study how well *in silico* interactions; correlated expression, shared transcription factors and genetic interactions serve to predict either type of interaction. Subsequently, we combine the information from in silico predictions, functional genomics data and protein interaction assays into evidence landscapes. In these landscapes we identify regions that are populated solely by physical or by metabolic interactions, allowing specific prediction of the nature of interactions between proteins.

Results

Performance of individual datasets on different reference sets

To distinguish metabolic interactions from physical interactions, we first investigate whether there are omics data that are typical for either of the two and therewith provide evidence for specific types of interactions. For various types of omics evidence we calculate the prediction accuracy for either physical or metabolic interactions (Figure 1), by comparing the predicted interactions with reference sets of either metabolic or physical interactions. Obviously, metabolic pathways contain multimeric enzyme complexes, but we do not score the intra-complex interaction of these as positives or as negatives in our metabolic reference. We do however consider the links between these enzymes and other enzymes from the pathway as metabolic interactions. Metabolic accuracy is then calculated as the total number of true metabolic interactions divided by the sum of the true and false metabolic and the true and false physical interactions. Physical accuracy is calculated as the total number of true physical interactions divided by the total number of true and false physical and the true and false metabolic interactions.

For each omics evidence type such as the correlation in the expression level of two genes, we calculate whether at a given score the interactions are metabolic, physical, or non-existent. The likelihood of physical interaction increases similarly to the likelihood of metabolic interactions for "simple" gene expression data, as well as for combinations of gene expression data between species or the combination of gene expression data with transcription factor binding data (ChIP-on-chip) (Figure 1, panel a to f). These data are therefore not specific for either metabolic or physical interactions. In contrast, for gene neighborhood the specific accuracy depends on the score: a very high level of gene neighborhood conservation is specific for physical interactions whereas a lower, but still significant level of gene neighborhood conservation is indicative of a metabolic interaction (Figure 1, panel g). The highest metabolic accuracy in this set is 0.83, at a neighborhood conservation score where the physical accuracy is 0.17. This dataset therefore contains some specificity about the type of interaction. Correlated phylogenetic profiles (panel h) show a similar, but less pronounced trend of differential specificity. Finally, we observe specificity for physical interactions not only in datasets where physical interaction was measured directly (panel j and k), but surprisingly also in one that contains the number of shared genetic interactions between proteins (panel i). For the protein complex purifications by Gavin (Gavin et al., 2006) and Krogan (Krogan et al., 2006) (panel j and k) it was of course expected that a high score in either of these sets is indicative for a physical interaction. Still, it is reassuring that these data are consistent with the metabolic interaction and physical interaction reference sets. Concluding, we can specifically predict physical interactions based on high quality protein-protein interaction screens and on shared Genetic Interactions. The ability to distinguish metabolic interactions is in fact the real challenge.

The evidence landscape: distinguishing metabolic from physical interactions.

Based on the observation that an intermediate score in gene neighborhood conservation is more indicative of a metabolic interaction than a high score (Figure 1g), we explore all pairwise combinations of high, intermediate and even null scores in pairs of genomics evidence types. We call these combinations of omics data "evidence landscapes", surfaces on which the x and y coordinates represent the scores of two types of genomics data, while the z coordinate represents a property of interest. We test the areas in these evidence landscapes for their specificity in reflecting either metabolic or physical interactions. In order to determine for a given region in the evidence landscape how well it predicts either type of interaction we define the differential accuracy. Differential accuracy is computed by subtracting the physical interaction accuracy from the metabolic interaction accuracy. This means that if a region scores equally well in both reference sets (be it very poor or very well) it has a zero differential accuracy, reflecting the inability of this region to differentiate between metabolic and physical interactions. However, if it is very accurate in predicting metabolic relations but unable to accurately predict physical interactions it has very high differential accuracy and, vice versa, a very negative value reflects specificity for physical interactions.

Figure 2 shows the differential accuracy in a representative selection of these evidence landscapes. The comprehensive collection of all evidence landscapes is available at our webpage (www.cmbi.ru.nl/~vvnoort/LANDSCAPE). Panel a shows the evidence landscape of the two TAP-TAG protein-protein interaction datasets (Gavin et al., 2006; Krogan et al., 2006).

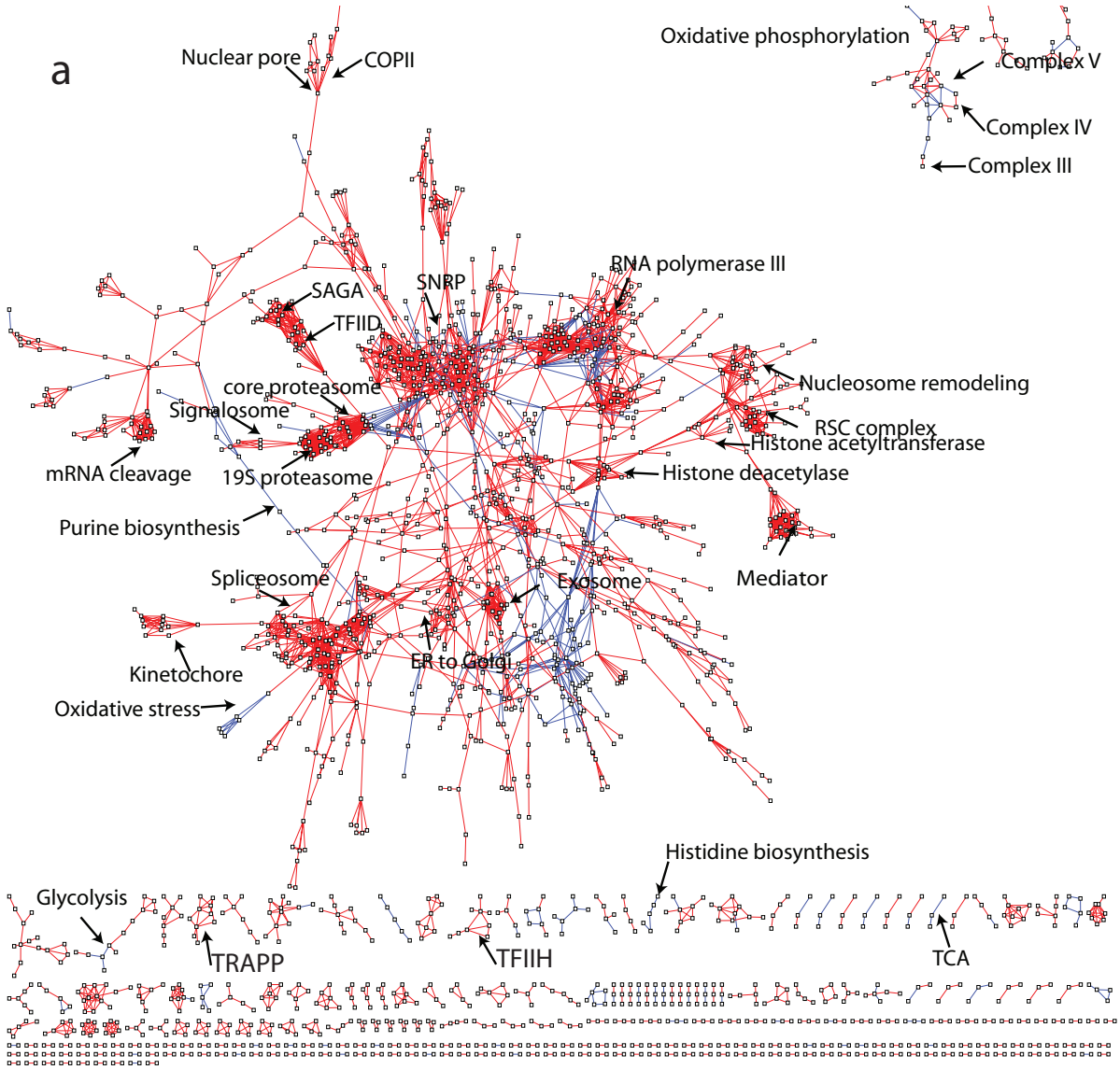
Despite the very high quality of both data sets, they are not completely comprehensive: each data set identifies interactions with a high Socio-affinity (Gavin et al., 2006) (SA) score between proteins that in the other assay were never co-purified, but which are true interactions in the physical interaction reference set. An SA score of 5 (bin 0.4) in only one of the two assays is not enough to predict a reliable physical interaction, however if the protein pair has an SA score of 5 in both sets it is a reliable prediction. So in fact the two assays complement each other. That is why in panel b and c we used the sum of the two SA scores for the evidence landscape. In these panels bin 0.2 contains all protein-pairs that were purified in both assays but never co-purified. Gene pairs with high orthologous conserved co-expression that were never co-purified are purely metabolic interactions (the upper left corner of Figure 2 panel b). We also observe this for co-regulated gene pairs (panel c). Indeed, we are now able to predict purely metabolic interactions by taking gene pairs that have a null score in the physical interaction set and positive in co-expression or co-regulation. Overlapping gene-pairs that are null scoring in the physical interaction datasets with gene pairs with an intermediate score of gene neighborhood or correlated phylogenetic profiles also yields purely metabolic interactions.

What we have observed in Figure 1 is that intermediate scores in correlated phylogenetic profiles and gene neighborhood conservation are relatively often indicative of metabolic interactions. The evidence landscape of these two has specific metabolic interactions in intermediate scores of both sets (panel g). Thus, not only do we find purely metabolic interactions from gene pairs that score null in protein-protein interaction datasets, we also find them in overlaps with intermediate scoring parts of other evidence types.

A cellular network that differentiates between physical interactions and functional associations

We extracted a list of predicted metabolic and physical interactions by taking all gene pairs from areas where the differential accuracy is either higher than 0.95 or lower than -0.95. This allows us to display a network of physical (red) and metabolic (blue) interactions (Figure 3a). Network visualizations are generally more open to biological interpretation than long lists of potential interactions. It is directly clear from the network layout that physical interactions are more clustered than metabolic interactions. The clustering coefficient (fraction of indirectly connected proteins that are also directly connected) of physical interactions (0.53) is much higher than the clustering coefficient of metabolic interactions (0.031). The incompleteness of the metabolic network relative to the physical interaction networks may bias this difference. However, the average number of connections per protein (K) is only twice as high for physical interactions (4.1) as for metabolic interactions (2.0) and the difference in clustering coefficients appears at least partly due to intrinsic difference between physical and metabolic interaction networks.

Several metabolic pathways are completely retrieved, like the Arginine and the Threonine biosynthesis pathways that are only connected by predicted metabolic interactions (blue lines). The Arginine biosynthesis pathway is depicted in Figure 3b. We find many known physical protein complexes as clusters densely connected by red lines as has been previously shown in many integrative bioinformatics studies (Han et al., 2004; Newman,



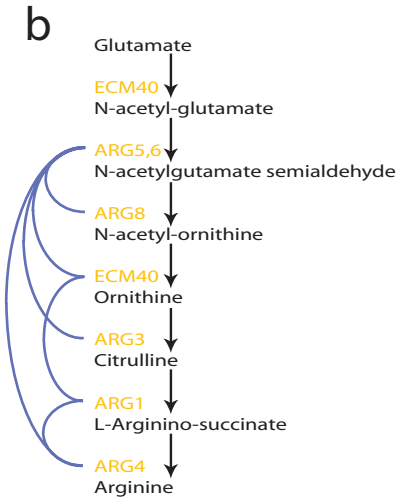


Figure 6.3 | Network of differentially predicted interactions. a The network of interactions in yeast that are specifically predicted to be physical (red lines) or metabolic (blue lines). We took all gene pairs that fell into squares (Figure 2) with a differential accuracy larger than 0.95 and at least five True Positive metabolic interactions for the specific metabolic interactions. We selected all gene pairs that fell into squares with differential accuracy smaller than -0.95 and at least five True Positive physical interactions for the specific physical interactions. Names of several known complexes and metabolic pathways are indicated on the network.
b The Arginine biosynthesis pathway in yeast. Names of the enzymes are in orange, arrows indicate biochemical reactions. Blue lines indicate all interactions that exist for these genes. Note that ECM40 catalyzes two steps in this pathway but the interactions with the other genes are drawn only once.

2006; Palla et al., 2005; Troyanskaya et al., 2003). Interestingly, we now also observe the pathway interactions that exist between them. For example, in the upper right corner is the oxidative phosphorylation pathway. Members of the same complex have red lines (physical interactions) between them, whereas members of different complexes have blue lines (metabolic interactions) between them. Even though we derived the metabolic pathway interactions by identifying the regions in the landscapes that scored high in a metabolic reference set we still expect this class in addition to be general for other functional associations from other types of cellular pathways. Therefore the blue lines between e.g. the exosome and the small nucleolar ribonucleoprotein complex are not necessarily metabolic like in the case of the oxidative phosphorylation pathway, but rather other types of functional associations. Likewise, the oxidative stress cluster contains interactions between thioredoxin reductases and glutaredoxins. These proteins are, as far as is known, not part of the same pathway in the sense that they pass e.g. reducing equivalents to each other, but they are part of the same system.

Discussion

When predicting interactions between genes it is essential to specify the type of interaction that is predicted to allow biological interpretation. Some data-types are already specific to the type of interaction, e.g. ChIP-on-chip data of transcription factors is indicative for regulatory interactions and co-purifications are specific for physical interactions. However, co-regulation, correlated expression, shared genetic interactions and in silico interactions are not intrinsically specific to any type of interaction. Here we have shown that although some datasets do contain a high level of metabolic interactions at intermediate scores, it is not possible to reliably predict metabolic interactions from them. However by combining the datasets in ways that examine the whole evidence landscape and not only the highest scoring protein pairs in both datasets, e.g. by taking protein-pairs that are evolutionarily conserved co-expressed but that never co-purify in two comprehensive protein-protein interaction datasets we can predict metabolic interactions.

It is perhaps logical in hindsight that we detect metabolic interactions in areas where both proteomic approaches report no co-purification while there are strong indications for co-regulation, but there are some important implications. We should not only use integrations based on the top scoring proteins but also use non-scoring proteins. For the co-purification data this implies that the absence of a reported interaction is in fact the reflection of a cellular reality: in other words we need physical protein interaction data sets where the negatives are really true negatives rather than the absence of results. Although the comparison of the Gavin and Krogan co-purification data reveals that both data sets still harbor some false negatives, a combined data set of both comes close to having the perfect properties for our objective, and it is only since the publication of these data that a differential genomics approach as proposed here has become possible.

Another contribution in distinguishing metabolic from physical interactions comes from differential rates of evolution. We could not obtain the same level of differential accuracy for the prediction of metabolic interactions in landscapes with the conserved co-expres-

sion set of Stuart and co-workers (2003) as with a two-species orthologous conserved co-expression (van Noort et al., 2003), because the first predicts mainly physical interactions. As the conserved co-expression set of Stuart is based on four species and the other one only on two, we speculate that metabolic interactions are less conserved in evolution than physical interactions, which is consistent with results on the evolutionary modularity of metabolic pathways and protein complexes in biological systems (Snel and Huynen, 2004). The higher rate of evolution of metabolic interactions also explains that a very high level of conservation of gene neighborhood conservation or correlation of phylogenetic profiles indicates a physical interaction whereas intermediate levels are more indicative of metabolic interactions.

It is of course tempting to combine more than two types of omics data. There are however two reasons why we here explore pairs of evidence types rather than to explore the multidimensional evidence landscape given by all evidence types simultaneously. Firstly, visual inspection of differential accuracy plots is still possible in two dimensions but becomes more troublesome in higher dimensions. Secondly, and more importantly, overlapping all evidence types at the same time results in very small numbers of protein pairs in each multidimensional volume in the reference sets, which in turn hampers the reliable calculation of prediction accuracy.

Protein relations predicted by our computational integration should be less laborious to experimentally test, because they prioritize the usability of various assays for biochemical verification. For example, it would be disingenuous to verify our metabolic relations by CoIP. In general, we expect that novel ways of integration and the advent of more and more types of omics data will allow the further development of approaches to increase the specificity and to extract more qualitative data on the nature of protein interactions.

Methods

Evidence types

Protein-protein interactions. We downloaded the yeast protein complex-purifications published by Gavin and co-workers (2006) and recalculated the Socio-affinity scores that reflect the likelihood of interaction to include also proteins that were purified only once. Protein pairs that are never co-purified but are both purified at least once get a socio-affinity score of zero. We also downloaded the protein complex-purifications of Krogan and co-workers (2006). These authors produced a different interaction score per protein pair, which was optimized to overlap with MIPS protein complexes. To have a reference set-independent score we calculated Socio-affinity scores based on the purifications of Krogan. Protein pairs that were never found together in a purification but were purified at least once were given a score of zero. As a third set we took the sum of Socio-affinity scores of all protein pairs occurring in both protein-protein interactions datasets.

In silico predictions of functional interactions were obtained from the database of STRING (von Mering et al., 2003). From this database we took the co-occurrence scores based on phylogenetic profiles of COGs and gene neighborhood conservation also based on

COGs. The scores were transferred from pairs of COGs to pairs of *S. cerevisiae* genes. If more than three yeast genes belonged to the same COG the score was considered ambiguous and was removed from the dataset.

Conserved co-expression. We used two multi-species conserved co-expression datasets; co-expression conservation between human, yeast, fly and worm (Stuart et al., 2003) and between yeast and worm (van Noort et al., 2003). We used also co-expression conservation between pairs of paralogs (van Noort et al., 2003) in yeast. For the two-species conservation we took the maximum expression correlation of all pairs of orthologs and averaged this maximum with the expression correlation of the gene-pair itself. For paralogous conservation we took the maximum expression correlation between all parallel duplicated gene pairs and averaged this maximum with the expression correlation of the gene-pair itself.

Co-regulation is assessed by combining correlated mRNA expression profiles with similarity in bound regulators to the gene promoter. Rick Young's lab made a comprehensive survey of the gene regulatory network in yeast (Harbison et al., 2004). We took a cut-off of 0.01 for binding of a transcription factor to a promoter based on the raw ChIP-on-chip data and divided the number shared transcription factors between two genes Ni,j by the geometric average of the total number of transcription factors bound by each of the two genes T resulting in co-regulation score Sij.

Gene pairs that share a promoter were excluded. To increase the reliability of the co-regulation signal, we multiply the correlation in binding profile by the correlation in mRNA expression profile based on a large-scale expression dataset in yeast (Stuart et al., 2003) i.e. Snew ij = rij *Sij where rij is the expression correlation of gene i and j.

Synthetic lethality. A set of synthetic lethal and synthetic sick interactions were downloaded from the Saccharomyces Genome Database (http://www.yeastgenome.org/). It was found earlier that genetic interactions (Tong et al., 2004) on their own are only marginally useful for predicting direct interactions, but shared genetic interactions do indicate involvement in similar pathways (Ozier et al., 2003). We corrected the number of shared genetic interactions Ni,j by the geometric average of total interactions T per protein exactly the same as the co-regulation score.

Reference sets

We downloaded known complexes from MIPS (Mewes et al., 2002) and removed all categories containing the terms 'other' or 'predicted'. We took complexes at the lowest level of definition. Protein pairs that are in the same complex are positive examples, protein pairs that are in different complexes are negative examples. The positive and negative examples constitute the physical interaction reference set.

From the KEGG database (Kanehisa et al., 2004) we took all metabolic maps with indices smaller than 2000. Maps with higher index are not metabolic and contain other processes including many that consist of a single protein complex. Positive examples are all protein pairs that co-occur on a metabolic map, negative examples are all protein pairs that do not co-occur on a metabolic map but are nevertheless present in the metabolic maps of KEGG. In order to not have any physical interactions in our metabolic reference set, we removed

all protein pairs with the same EC number and removed all protein pairs that are part of the same complex according to SGD/GO annotation (Ashburner et al., 2000; Dwight et al., 2002) or MIPS. Together the positive and negative examples form the metabolic interaction reference set.

Cytoplasmic ribosomal proteins were removed from all reference sets and datasets. As they confer very many pair-wise interactions, including them would bias the statistics towards ribosomes.

Accuracy and differential accuracy

The conserved co-expression values of the Kim lab (Stuart et al., 2003) were rescaled by transforming the -log(P value) to scores between 0 and 1, such that high scores correspond to more likely interactions. All other scores were rescaled to scores between 0 and 1 by a linear transformation. In the score-accuracy plots for each set a binning was made with bin width 0.025, bins containing fewer than five gene pairs were added to the preceding bin to avoid small number statistics. In the evidence landscape plots, we plot two data sets against each other in a heat map like fashion and color squares according to their differential accuracy (see below). Squares were made with sides of 0.05, if a square contained fewer than two True Positives a larger square with sides 0.1 was made to avoid high accuracies based on very few examples.

Physical interaction accuracy (A *phys*) was calculated as the number of True Positives of the physical interaction reference set divided by the number of True Positives plus False Positives of both reference set sets in that bin. Metabolic interaction (A *meta*) accuracy was calculated as the number of True Positives of the metabolic interaction reference set divided by the number of True Positives plus False Positives of both reference set sets in that bin. Total accuracy was calculated as the number of True Positives divided by the number of True Positives plus False Positives of both reference set sets in that bin. In order to score for a given region/square bin in the evidence landscape how well it predicts either type of interaction we compute what we here call the differential accuracy (A *diff*).

	Positive metabolic	Negative metabolic	Positive physical	Negative physical
Present in bin	TP *meta*	FP *meta*	TP *phys*	FP *phys*

A *meta* = TP *meta* / (TP *meta* + FP *meta* + TP *phys* + FP *phys*)
A *phys* = TP *phys* / (TP *meta* + FP *meta* + TP *phys* + FP *phys*)
A *total* = A *meta* + A *phys*
A *diff* = A *meta* – A *phys*

Differential accuracy is computed by subtracting the physical interaction accuracy from the metabolic interaction accuracy. This means that if a region scores equally well in both reference sets (be it very poor or very well) it has a zero differential accuracy reflecting the inability of this region to differentiate between metabolic and physical interactions. However, if it is very accurate in predicting metabolic relations but unable to accurately predict physical interactions is has very high differential accuracy and vice versa a very negative value reflects specificity for physical interactions.

Adding specificity to predicted interactions

We took all gene pairs that fell into squares with differential accuracy larger than 0.95 and at least five True Positive metabolic interactions and called them *predicted metabolic interactions*. We selected all gene pairs that fell into squares with differential accuracy smaller than -0.95 and at least five True Positive physical interactions and called them *predicted physical interactions*.

Software

Figure 1 was made using xmgrace (http://plasma-gate.wizmann.ac.il/Grace). The panels of Figure 2 were made using R (www.R-project.org). The network of predicted interactions is visualized using cytoscape (www.cytoscape.org).

Chapter 7

Summarizing discussion

Summarizing discussion

This thesis focuses on developing comparative genomics methods in eukaryotes, with an emphasis on applications for gene function prediction and regulatory element detection. In this chapter, I will summarize the main lessons that we have learned from our results and provide an outlook on future possibilities of comparative genomics.

Functional associations in eukaryotes

In the past, methods have been developed to predict functional associations between gene pairs in prokaryotes (Huynen and Bork, 1998; Marcotte et al., 1999; Overbeek et al., 1999; Pellegrini et al., 1999; Snel et al., 2000). Three different methods exist, gene neighborhood conservation, gene fusion and gene co-occurrence. The development of these methods has been facilitated by a number of conditions. Firstly, the operon structure of prokaryotic messenger RNAs makes it easy to identify co-expressed genes directly from the genome sequence by finding genes that are in each others neighborhood. Secondly, a large number of prokaryotic genomes have been sequenced. This has made it possible to identify which genes are potentially co-expressed throughout a number of organisms (conserved gene-neighborhood), to identify genes that are present and absent in the same organisms (co-occurrence) and to identify genes that are fused in some organisms but not in others (gene fusion). Finally, most gene families tend to duplicate less often in prokaryotes than in eukaryotes, making the definition of orthology easier in prokaryotes than in eukaryotes where duplications are ubiquitous.

The challenge of this thesis was to extend the genomic association methods to eukaryotes. In the absence of operon structure in eukaryotes, we need a different way to find co-expressed genes. In two eukaryotic organisms, yeast and worm, large-scale mRNA expression measurements had been performed using genome chips (Hughes et al., 2000; Kim et al., 2001). Co-expressed genes can be found by calculating the correlations between the expression profiles of individual genes. One challenge in finding evolutionary conserved co-expression is to define orthologs. For this, we developed a method based on phylogenetic trees, that includes the possibility of multiple orthologous relations per gene. The phylogenetic tree method cannot easily be extended to more than two species, but recently, we have developed a method that is able to detect groups of orthologs in multiple species based on phylogenetic trees (van der Heijden et al., 2006). In the future, this method could be used for finding evolutionary conservation of interactions in eukaryotes. We have shown that gene pairs that have conserved co-expression are much more likely to act in the same pathway than random gene pairs or gene pairs that are co-expressed in only one species. In parallel to our work, another method of finding evolutionary conservation of co-expression was

developed by Stuart and co-workers (Stuart et al., 2003). They used Bidirectional Best Hits as orthology definition which yields a much smaller number of yeast genes for which one can potentially find conserved co-expression than the phylogenetic tree method. Where we applied a threshold to the expression correlation, they predicted interactions based on the order of gene expression correlations to a query gene. A gene is predicted to interact with the query gene if it occurs as top-scoring in the ordered gene lists of four organisms. The differences in methodology also result in differences in predicted interactions. Our predicted interactions are a mixture of physical and metabolic interactions, whereas the interactions predicted by Stuart and co-workers mainly of physical interactions. We also applied conservation of interactions to yeast-2-hybrid assays in yeast and fly (Giot et al., 2003; Ito et al., 2001; Uetz et al., 2000) and revealed a significant level of conservation of physical interactions. Again, the conserved interactions proved to be more reliable than the non-conserved interactions. In analogy to horizontal gene transfer, which is the transfer of genes between species, we call these methods horizontal comparative genomics.

A second method of evolutionary conservation takes advantage of the large number of gene duplications in eukaryotes. We start with identifying pairs of genes (A and B) that are co-expressed. Then, we investigate whether a pair of paralogous genes exists (A' and B') that are also co-expressed. We found that pairs of paralogous genes with conserved co-expression are also likely to act in the same pathway. Like we have shown for gene co-expression, Deane and co-workers (Deane et al., 2002) have shown that physical interactions that are conserved after parallel gene duplication are also more reliable than unconserved interactions.

A third method to predict functional associations takes the overlap of predicted interactions from multiple data types as was first suggested by von Mering and co-workers (2002). Expression data on the mRNA and protein level have become available from the malaria parasite *Plasmodium falciparum* (Bozdech et al., 2003; Florens et al., 2002; Lasonder et al., 2002; Le Roch et al., 2003). We constructed a full Bayesian network that predicts interactions between genes based on co-expression. If an interaction is predicted by co-expression in three out of four datasets, it is highly likely to be a real interaction. The "conservation" of co-expression is now not defined between pairs of paralogous or orthologous genes but between functional genomics assays. We coined the term vertical comparative genomics to summarize methods that determine conservation between different data types within a species. As no orthology or paralogy definitions are necessary for this method, it can easily be applied to organisms where multiple sources of functional genomics data are available. Another vertical comparative genomics effort was made by constructing a naïve Bayesian network to predict physical protein interactions, effectively adding up the log odds probabilities of interaction from heterogeneous data sources (Troyanskaya et al., 2003). A combination of a naïve and a full Bayesian network is even better at predicting physical interactions (Jansen et al., 2003). A third type of integration is made in the STRING database (von Mering et al., 2003), where one final probability of an interaction between two genes is calculated by multiplying the probabilities that the genes are not interacting as predicted by various data sources.

Summarizing, in this thesis, parallel to other groups, we have developed three different types of methods that can predict functional associations between gene pairs in eukaryotes;

conservation of interactions between orthologous gene pairs, conservation of interactions between paralogous gene pairs and conservation of interactions between different data sources from the same species.

Adding specificity to functional associations

Although comparative genomics has been successful in increasing the reliability of protein-protein interactions that can be predicted from genomics data, it gives no information about the type of interaction that is expected to be found. This lack of specificity can be a hurdle when using the functional associations to predict functions for individual proteins and to devise experiments to validate the interactions. We have shown that some methods for finding functional associations are specific to physical interactions whereas others predict a mixture of interactions including physical interactions and metabolic interactions. Instead of collapsing all predictions into one final score, we have found combinations of scores that are specific for either metabolic or physical interactions. The most straightforward combinations are gene pairs that score high in a dataset that is not specific like conserved co-expression and low in a physical interaction assay. Other, less intuitive combinations are gene pairs with intermediate scores in two datasets that in themselves are not specific for one type of interaction, like gene neighborhood conservation and co-occurrence. Both combinations retrieve interactions that are metabolic rather than physical. Thus, by taking only these predicted metabolic interactions we are able to retrieve complete metabolic pathways. Using such differential genomics methods, we are able to in parallel retrieve physical interactions and metabolic interactions. For example, we retrieved the physical interactions that exist within the complexes of the oxidative phosphorylation, as well as the metabolic interactions that exist between these five complexes.

Regulatory element detection

The simultaneous expression of genes is the result of regulation by the same transcription factors amongst others. The binding of these transcription factors is determined by a specific DNA sequence. Two distinct methods exist to detect these elements computationally. The first relies on the conservation of functional elements in multiple species. If a certain DNA sequence has some function to regulate the expression of the downstream gene, a mutation would harm this function and thus the DNA sequence is expected to be conserved. Phylogenetic footprinting (Blanchette and Tompa, 2002) is the method that finds conserved sequences of DNA between upstream regions of orthologous genes and is named after its experimental equivalent to find the location of binding of regulatory proteins. The second method defines groups of co-expressed genes and finds DNA sequences that are overrepresented in the upstream regions of these genes but not in other groups (Roth et al., 1998; van Helden et al., 1998). We have synthesized these two methods to find DNA elements that are overrepresented in groups of co-expressed genes in *Plasmodium falciparum* as well as their orthologs *Plasmodium yoelii yoelii*. We show that this method is more successful in detect-

ing elements that correlate with mRNA expression than using just the co-expressed genes. An unexpected outcome of this research was that we found much more regulatory elements per upstream region than in Saccharomyces cerevisiae; a eukaryote with a similar number of genes. In conjunction with the small number of specific transcriptional regulators, this suggested that *Plasmodium falciparum* uses combinations of regulators to obtain diversity in expression patterns.

Network evolution

Graph theoreticians have shown that biological networks like many other networks often have a non-random topology. This means that distributing the interactions between genes randomly over all gene pairs would result in a very different topology than the network topology that we observe, e.g. in metabolic networks. It has lead to the hypothesis that there may be selection pressure acting on the overall network topology (Fell and Wagner, 2000; Jeong et al., 2001). We found that the *S. cerevisiae* co-expression network also had this non-random topology; it had a scale-free distribution of number of links per gene (Barabasi and Albert, 1999) and showed the 'small-world' effect (Watts and Strogatz, 1998). However, biomolecular interactions are not distributed randomly but are inherited from the parent organism and may have some random changes due to mutations. Specifically, co-expression is a result of regulation by the same transcription factors and transcription factor binding sites that are inherited. We modeled the process of neutral evolution of gene regulation and derived 'random' co-expression networks. A wide class of parameter settings resulted in network topologies very similar to the real co-expression network. By parsimonious arguments, we conclude that there is no selection pressure acting on the global network topology of the co-expression network. A more general conclusion is that phenomena the we observe in biology that seem non-random do not have to be favored by selection but may simply be the outcome of neutral evolution and the mechanisms of gene duplication.

Outlook

Comparative genomics has moved from prokaryotes to eukaryotes and has therewith allowed the prediction of gene regulatory elements and specific functional associations in these species. More and more functional genomics data and genome sequences have become and will become available. Not only more data, but also more different types of data will become available as more techniques are developed. Most of the methods are established in yeast, but will be extended to other organisms and likely the quality of data will increase as biologists are getting more experienced with the techniques. The simple overlap of functional associations between and within organisms has proven to be useful to make the predicted associations more reliable. However, the ever increasing amounts of data allow and demand more creative ways of integration in the form of differential genomics to get reliability of predicted interactions in combination with specificity on the type of interactions.

A pitfall of the post-genomic era, where it is possible to measure everything at the

same time is that many genome-scale experiments are devised without a specific question in mind. With the sequencing of genomes, biologists have become a bit like stamp collectors and collect as much experimental data on their biological system as they can. In the end, the purpose of collecting all these data is the gain of biological understanding of the biological systems, but how this understanding can be reached is often not straightforward. Comparative genomics has the same pitfall; although data integration can be useful, we should not just integrate everything with everything without a specific purpose. We should keep in the back of our minds or preferably in the front of our minds that the integration of genomics, functional genomics and proteomics data should lead to more understanding of the biology of the cell and the complete organism.

Within comparative genomics, this biological insight comes on the one hand from the discovery of gene functions and interactions, and on the other hand from studying the evolution of the interactions between genes. In this thesis, we have also paid attention to the aspect of evolutionary processes. For example, we found that there is significant conservation of both co-expression and physical interactions between organisms and that networks that appear favored by selection can simply come about by neutral evolution. Still, much more can be learned. A study on modularity of biomolecular systems has shown that metabolic interactions are less well conserved than physical interactions (Snel and Huynen, 2004). A question that remains is how the individual subunits of complexes and members of metabolic pathways differ in their evolutionary stability. Can we identify core subunits that are always present and others that are more variable in their presence-absence patterns? Another remaining question is how metabolic pathways, protein complexes and other biological systems evolve with respect to each other. We may be able to draw a parallel with co-occurrence of individual genes and study which pathways and complexes also tend to be absent and present together in different organisms. The insight that will be gained by studying the evolutionary processes will enhance the development of new comparative genomics methods for the reconstruction of complete biological systems.

Co-expression, physical interactions and regulatory interactions between pathways may also differ from one organism to the other, depending on the life-style or requirements of the organism. They may even differ within one organism, depending on environmental conditions, developmental stage or cell-type. Differential genomics may provide new insight into the variation that is present between different organisms not only on the level of the individual gene but also on the level of whole pathways and complexes. In the near future, we will be able to study the presence and absence of these biological systems in time and in space as more and more techniques will be developed and applied in more different organisms. At the same time computational methods will have to be developed to analyze, interpret and integrate all these new data that will not only consist of numerical data (like expression levels), but also more and more of visual data (like movies and photographs). Mass spectrometry methods will also produce more and more data concerning the levels of protein expression as well as levels of metabolites. The field of comparative genomics will gain a whole new dimension by the integration of metabolomics with functional genomics and proteomics and potentially cell biology data that comes in the form of pictures and movies.

Bibliography

(1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. Science, 282, 2012-2018.
Aloy, P. and Russell, R.B. (2002) Potential artefacts in protein-interaction networks. FEBS Lett, 530, 253-254.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J Mol Biol, 215, 403-410.

Amaral, L.A., Scala, A., Barthelemy, M. and Stanley, H.E. (2000) Classes of small-world networks. Proc Natl Acad Sci U S A, 97, 11149-11152.

Anderson, N.L. and Anderson, N.G. (1998) Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis, 19, 1853-1861.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25, 25-29.

Bahl, A., Brunk, B., Crabtree, J., Fraunholz, M.J., Gajria, B., Grant, G.R., Ginsburg, H., Gupta, D., Kissinger, J.C., Labo, P., Li, L., Mailman, M.D., Milgram, A.J., Pearson, D.S., Roos, D.S., Schug, J., Stoeckert, C.J., Jr. and Whetzel, P. (2003) PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. Nucleic Acids Res, 31, 212-215.

Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks. Science, 286, 509-512.

Beyer, A., Workman, C., Hollunder, J., Radke, D., Moller, U., Wilhelm, T. and Ideker, T. (2006) Integrated assessment and prediction of transcription factor binding. PLoS Comput Biol, 2, e70.

Bhan, A., Galas, D.J. and Dewey, T.G. (2002) A duplication growth model of gene expression networks. Bioinformatics, 18, 1486-1493.

Blanchette, M. and Tompa, M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res, 12, 739-748.

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J. and DeRisi, J.L. (2003) The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. PLoS Biol, 1, E5.

Brummelkamp, T.R. and Bernards, R. (2003) New tools for functional mammalian cancer genetics. Nat Rev Cancer, 3, 781-789.

Bussemaker, H.J., Li, H. and Siggia, E.D. (2001) Regulatory element detection using correlation with expression. Nat Genet, 27, 167-171.

Calderwood, M.S., Gannoun-Zaki, L., Wellems, T.E. and Deitsch, K.W. (2003) Plasmodium falciparum var genes are regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron. J Biol Chem, 278, 34125-34132.

Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., et al. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature, 419, 512-519.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A. and Johnston, M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science, 301, 71-76.

Coombs, G.H., Goldberg, D.E., Klemba, M., Berry, C., Kay, J. and Mottram, J.C. (2001) Aspartic proteases of Plasmodium falciparum and other parasitic protozoa as drug targets. Trends Parasitol, 17, 532-537.

Coulson, R.M., Hall, N. and Ouzounis, C.A. (2004) Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum. Genome Res, 14, 1548-1554.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. Genome Res, 14, 1188-1190.

Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci, 23, 324-328.

Davidson, E.H., McClay, D.R. and Hood, L. (2003) Regulatory gene networks and the properties of the developmental process. Proc Natl Acad Sci U S A, 100, 1475-1480.

De Virgilio, C., Burckert, N., Bell, W., Jeno, P., Boller, T. and Wiemken, A. (1993) Disruption of TPS2, the gene encoding the 100-kDa subunit of the trehalose-6-phosphate synthase/phosphatase complex in Saccharomyces cerevisiae, causes accumulation of trehalose-6-phosphate and loss of trehalose-6-phosphate phosphatase activity. Eur J Biochem, 212, 315-323.

Deane, C.M., Salwinski, L., Xenarios, I. and Eisenberg, D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics, 1, 349-356.

Dechering, K.J., Kaan, A.M., Mbacham, W., Wirth, D.F., Eling, W., Konings, R.N. and Stunnenberg, H.G. (1999) Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite Plasmodium falciparum. Mol Cell Biol, 19, 967-978.

Deitsch, K.W., Calderwood, M.S. and Wellems, T.E. (2001) Malaria. Cooperative silencing elements in var genes. Nature, 412, 875-876.

Dwight, S.S., Harris, M.A., Dolinski, K., Ball, C.A., Binkley, G., Christie, K.R., Fisk, D.G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D. and Cherry, J.M. (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res, 30, 69-72.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A, 95, 14863-14868.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Protein interaction

maps for complete genomes based on gene fusion events. Nature, 402, 86-90.
Fang, J. and McCutchan, T.F. (2002) Thermoregulation in a parasite's life cycle. Nature, 418, 742.
Fang, J., Sullivan, M. and McCutchan, T.F. (2004) The effects of glucose concentration on the reciprocal regulation of rRNA promoters in Plasmodium falciparum. J Biol Chem, 279, 720-725.
Fell, D.A. and Wagner, A. (2000) The small world of metabolism. Nat Biotechnol, 18, 1121-1122.
Fields, S., Kohara, Y. and Lockhart, D.J. (1999) Functional genomics. Proc Natl Acad Sci USA, 96, 8825-8826.
Fields, S. and Song, O. (1989) A novel genetic system to detect protein-protein interactions. Nature, 340, 245-246.
Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. Syst Zool, 19, 99-113.
Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. and et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269, 496-512.
Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L., Witney, A.A., Wolters, D., Wu, Y., Gardner, M.J., Holder, A.A., Sinden, R.E., Yates, J.R. and Carucci, D.J. (2002) A proteomic view of the Plasmodium falciparum life cycle. Nature, 419, 520-526.
Forster, J., Famili, I., Palsson, B.O. and Nielsen, J. (2003) Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. Omics, 7, 193-202.
Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. and Feldman, M.W. (2002) Evolutionary rate in the protein interaction network. Science, 296, 750-752.
Galburt, E., Pelletier, J., Wilson, G. and Stoddard, B. (2002) Structure of a tRNA Repair Enzyme and Molecular Biology Workhorse. T4 Polynucleotide Kinase. Structure (Camb), 10, 1249.
Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature, 440, 631-636.
Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415, 141-147.
Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K. and Weissman, J.S. (2003) Global analysis of protein expression in yeast. Nature, 425, 737-741.
Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., et al. (2003) A protein interaction map of Drosophila melanogaster. Science, 302, 1727-1736.
Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996) Life with 6000 genes. Science, 274, 546, 563-547.

Grandori, R., Khalifah, P., Boice, J.A., Fairman, R., Giovanielli, K. and Carey, J. (1998) Biochemical characterization of WrbA, founding member of a new family of multimeric flavodoxin-like proteins. J Biol Chem, 273, 20960-20966.

Guelzim, N., Bottani, S., Bourgine, P. and Kepes, F. (2002) Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet, 31, 60-63.

Hall, N., Karras, M., Raine, J.D., Carlton, J.M., Kooij, T.W., et al. (2005) A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. Science, 307, 82-86.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P. and Vidal, M. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 430, 88-93.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E. and Young, R.A. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature, 431, 99-104.

Hayano, K. and Fukui, S. (1967) Purification and properties of 3-ketosucrose-forming enzyme from the cells of Agrobacterium tumefaciens. J Biol Chem, 242, 3655-3672.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature, 415, 180-183.

Horrocks, P. and Lanzer, M. (1999) Mutational analysis identifies a five base pair cis-acting sequence essential for GBP130 promoter activity in Plasmodium falciparum. Mol Biochem Parasitol, 99, 77-87.

Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000a) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol, 296, 1205-1214.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. and Friend, S.H. (2000b) Functional discovery via a compendium of expression profiles. Cell, 102, 109-126.

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S. and O'Shea, E.K. (2003) Global analysis of protein localization in budding yeast. Nature, 425, 686-691.

Huynen, M., Snel, B., Lathe, W., 3rd and Bork, P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res, 10, 1204-1210.

Huynen, M.A. and Bork, P. (1998) Measuring genome evolution. Proc Natl Acad Sci U S A, 95, 5849-5856.

Huynen, M.A., Snel, B., Bork, P. and Gibson, T.J. (2001) The phylogenetic distribution of frataxin indicates a role in iron-sulfur cluster protein assembly. Hum Mol Genet, 10, 2463-2468.

Huynen, M.A., Snel, B. and van Noort, V. (2004) Comparative genomics for reliable protein-function prediction from genomic data. Trends Genet, 20, 340-344.
Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A, 98, 4569-4574.
Jansen, R., Greenbaum, D. and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. Genome Res, 12, 37-46.
Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 302, 449-453.
Jensen, L.J., Lagarde, J., von Mering, C. and Bork, P. (2004) ArrayProspector: a web resource of functional associations inferred from microarray expression data. Nucleic Acids Res, 32, W445-448.
Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. Nature, 411, 41-42.
Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) The large-scale organization of metabolic networks. Nature, 407, 651-654.
Joyce, A.R. and Palsson, B.O. (2006) The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol, 7, 198-210.
Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res, 32, D277-280.
Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R. and Ideker, T. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci U S A, 100, 11394-11399.
Kelley, R. and Ideker, T. (2005) Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol, 23, 561-566.
Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N. and Davidson, G.S. (2001) A gene expression map for Caenorhabditis elegans. Science, 293, 2087-2092.
Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature, 440, 637-643.
LaCount, D.J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J.R., Schoenfeld, L.W., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S. and Hughes, R.E. (2005) A protein interaction network of the malaria parasite Plasmodium falciparum. Nature, 438, 103-107.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., et al. (2001) Initial sequencing and analysis of the human genome. Nature, 409, 860-921.
Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M., Pain, A., Sauerwein, R.W., Eling, W.M., Hall, N., Waters, A.P., Stunnenberg, H.G. and Mann, M. (2002) Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. Nature, 419, 537-542.
Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J. and Winzeler, E.A. (2003) Discovery of gene

function by expression profiling of the malaria parasite life cycle. Science, 301, 1503-1508.
Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. and Young, R.A. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science, 298, 799-804.
Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., et al. (2004) A map of the interactome network of the metazoan C. elegans. Science, 303, 540-543.
Liu, Y., Liu, X.S., Wei, L., Altman, R.B. and Batzoglou, S. (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. Genome Res, 14, 451-458.
Ma, H. and Zeng, A.P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics, 19, 270-277.
Marbois, B.N. and Clarke, C.F. (1996) The COQ7 gene encodes a protein in Saccharomyces cerevisiae necessary for ubiquinone biosynthesis. J Biol Chem, 271, 2995-3004.
Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences. Science, 285, 751-753.
Maslov, S. and Sneppen, K. (2002) Specificity and stability in topology of protein networks. Science, 296, 910-913.
Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B. (2002) MIPS: a database for genomes and protein sequences. Nucleic Acids Res, 30, 31-34.
Militello, K.T., Dodge, M., Bethke, L. and Wirth, D.F. (2004) Identification of regulatory elements in the Plasmodium falciparum genome. Mol Biochem Parasitol, 134, 75-88.
Mushegian, A.R. and Koonin, E.V. (1996) Gene order is not conserved in bacterial evolution. Trends Genet, 12, 289-290.
Newman, M.E. (2006) Modularity and community structure in networks. Proc Natl Acad Sci U S A, 103, 8577-8582.
Nitta, M., Saijo, M., Kodo, N., Matsuda, T., Nakatsu, Y., Tamai, H. and Tanaka, K. (2000) A novel cytoplasmic GTPase XAB1 interacts with DNA repair protein XPA. Nucleic Acids Res, 28, 4212-4218.
Noordewier, M.O. and Warren, P.V. (2001) Gene expression microarrays and the integration of biological knowledge. Trends Biotechnol, 19, 412-415.
O'Brien, K.P., Remm, M. and Sonnhammer, E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res, 33, D476-480.
Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res, 27, 29-34.
Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N. (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A, 96, 2896-2901.
Ozier, O., Amin, N. and Ideker, T. (2003) Global architecture of genetic interactions on the protein network. Nat Biotechnol, 21, 490-491.

Palla, G., Derenyi, I., Farkas, I. and Vicsek, T. (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435, 814-818.
Pastor-Satorras, R., Smith, E. and Sole, R.V. (2003) Evolving protein interaction networks through gene duplication. J Theor Biol, 222, 199-210.
Pearson, W.R. (1998) Empirical statistical estimates for sequence similarity searches. J Mol Biol, 276, 71-84.
Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A, 96, 4285-4288.
Peng, W.T., Robinson, M.D., Mnaimneh, S., Krogan, N.J., Cagney, G., et al. (2003) A panoramic view of yeast noncoding RNA processing. Cell, 113, 919-933.
Pereira-Leal, J.B., Enright, A.J. and Ouzounis, C.A. (2004) Detection of functional modules from protein interaction networks. Proteins, 54, 49-57.
Polson, H.E. and Blackman, M.J. (2005) A role for poly(dA)poly(dT) tracts in directing activity of the Plasmodium falciparum calmodulin gene promoter. Mol Biochem Parasitol, 141, 179-189.
Porter, M.E. (2002) Positive and negative effects of deletions and mutations within the 5' flanking sequences of Plasmodium falciparum DNA polymerase delta. Mol Biochem Parasitol, 122, 9-19.
Purnelle, B., Skala, J., van Dyck, L. and Goffeau, A. (1994) Analysis of an 11.7 kb DNA fragment of chromosome XI reveals a new tRNA gene and four new open reading frames including a leucine zipper protein and a homologue to the yeast mitochondrial regulator ABF2. Yeast, 10, 125-130.
Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A. and Legrain, P. (2001) The protein-protein interaction map of Helicobacter pylori. Nature, 409, 211-215.
Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabasi, A.L. (2002) Hierarchical organization of modularity in metabolic networks. Science, 297, 1551-1555.
Rea, S. (2001) CLK-1/Coq7p is a DMQ mono-oxygenase and a new member of the di-iron carboxylate protein family. FEBS Lett, 509, 389-394.
Remm, M., Storm, C.E. and Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol, 314, 1041-1052.
Richmond, C.S., Glasner, J.D., Mau, R., Jin, H. and Blattner, F.R. (1999) Genome-wide expression profiling in Escherichia coli K-12. Nucleic Acids Res, 27, 3821-3835.
Roth, F.P., Hughes, J.D., Estep, P.W. and Church, G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol, 16, 939-945.
Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol, 4, 406-425.
Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270, 467-470.
Simon, H.A. and Bonini, C.P. (1958) The size distribution of business firms. American Economic Review, 48, 607-617.

(in Saccharomyces cerevisiae). Proc Natl Acad Sci U S A, 100, 8348-8353.
Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.M. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 403, 623-627.
van Beeumen, J. and de Ley, J. (1975) A ferrodoxin from Agrobacterium tumefaciens. FEBS Lett, 59, 146-148.
van der Heijden, R.T.J.M., Snel, B., van Noort, V. and Huynen, M.A. (submitted) Orthology prediction at Scalable resolution by Phylogenetic Tree analysis.
van Helden, J., Andre, B. and Collado-Vides, J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol, 281, 827-842.
van Noort, V., Snel, B. and Huynen, M.A. (2003) Predicting gene function by conserved co-expression. Trends Genet, 19, 238-242.
Vance, J.R. and Wilson, T.E. (2001) Repair of DNA strand breaks by the overlapping functions of lesion-specific and non-lesion-specific DNA 3' phosphatases. Mol Cell Biol, 21, 7191-7198.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., et al. (2001) The sequence of the human genome. Science, 291, 1304-1351.
von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P. and Snel, B. (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res, 31, 258-261.
von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 417, 399-403.
Wagner, A. (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proc Natl Acad Sci U S A, 97, 6579-6584.
Wagner, A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol, 18, 1283-1292.
Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. Nature, 393, 440-442.
Wu, L.F., Hughes, T.R., Davierwala, A.P., Robinson, M.D., Stoughton, R. and Altschuler, S.J. (2002) Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. Nat Genet, 31, 255-265.
Wuchty, S., Oltvai, Z.N. and Barabasi, A.L. (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat Genet, 35, 176-179.
Yanai, I. and DeLisi, C. (2002) The society of genes: networks of functional links between genes from comparative genomics. Genome Biol, 3, research0064.
Yen, P.H., Ellison, J., Salido, E.C., Mohandas, T. and Shapiro, L. (1992) Isolation of a new gene from the distal short arm of the human X chromosome that escapes X-inactivation. Hum Mol Genet, 1, 47-52.

Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. J Mol Biol, 147, 195-197.
Snel, B., Bork, P. and Huynen, M.A. (2002) The identification of functional modules from the genomic association of genes. Proc Natl Acad Sci U S A, 99, 5890-5895.
Snel, B. and Huynen, M.A. (2004) Quantifying modularity in the evolution of biomolecular systems. Genome Res, 14, 391-397.
Snel, B., Lehmann, G., Bork, P. and Huynen, M.A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res, 28, 3442-3444.
Snel, B., van Noort, V. and Huynen, M.A. (2004) Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. Nucleic Acids Res, 32, 4725-4731.
Sonoda, E., Okada, T., Zhao, G.Y., Tateishi, S., Araki, K., Yamaizumi, M., Yagi, T., Verkaik, N.S., van Gent, D.C., Takata, M. and Takeda, S. (2003) Multiple roles of Rev3, the catalytic subunit of polzeta in maintaining genome stability in vertebrates. Embo J, 22, 3188-3197.
Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell, 9, 3273-3297.
Spirin, V. and Mirny, L.A. (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A, 100, 12123-12128.
Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. and Gilles, E.D. (2002) Metabolic network structure determines key aspects of functionality and regulation. Nature, 420, 190-193.
Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science, 302, 249-255.
Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res, 29, 22-28.
Teichmann, S. and Babu, M. (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. Trends Biotechnol, 20, 407.
Teichmann, S.A., Rison, S.C., Thornton, J.M., Riley, M., Gough, J. and Chothia, C. (2001) The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. J Mol Biol, 311, 693-708.
Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 22, 4673-4680.
Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., et al. (2004) Global mapping of the yeast genetic interaction network. Science, 303, 808-813.
Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B. and Botstein, D. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction

Yu, H., Zhu, X., Greenbaum, D., Karro, J. and Gerstein, M. (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. Nucleic Acids Res, 32, 328-337.

Samenvatting voor iedereen

Voor de meeste van jullie is wat ik de afgelopen jaren heb gedaan niet veel duidelijker geworden dan 'iets met genen'. In dit hoofdstuk wil ik voor iedereen wat duidelijker maken wat dat 'iets' nou eigenlijk inhield. De verzameling van alle genen van een organisme, noemen we het genoom. In de afgelopen tien jaar hebben biologen de complete genomen van zowel kleine parasieten als multicellulaire organismen zoals wijzelf in kaart gebracht. Een genoom bestaat uit hele grote moleculen DNA die we in blokjes onder kunnen verdelen. We hebben steeds een blokje fosfaat en een blokje suiker met daaraan een nucleotide. Dit patroon herhaalt zich miljoenen malen. Van de nucleotiden hebben we er vier. Binnen een gen (dat is dus een gedeelte van het genoom) bepaalt de specifieke volgorde van deze nucleotiden welke eiwitten worden gevormd. De eiwitten zorgen er voor dat alle biochemische reacties in de cel plaatsvinden. Met het genoom zouden we dus een blauwdruk van een organisme in handen hebben en de verwachtingen van de genoom projecten waren dan ook hoog. Het bleek echter dat het kennen van de volgorde van de nucleotiden in de genomen, op zichzelf niet voldoende was om te begrijpen wat er zich binnen in de biologische cel allemaal afspeelde. Dit komt onder andere doordat we eigenlijk geen helder beeld hebben over hoe de genen samenwerken om het complete organisme te maken. Daarom zijn biologen begonnen met het op grote schaal verzamelen van informatie, zoals onder welke omstandigheden die genen eigenlijk afgeschreven worden om er eiwitten van te maken. Deze grootschalige data noemen we genomics data, omdat ze informatie geven over het hele genoom.

Een voorbeeld van genomics data is mRNA expressie. Een mRNA molecuul is een kopie van een gen, die de cel gebruikt om eiwitten te produceren. Vergelijk het met de bouw, een aannemer gaat niet met de blauwdrukken naar de bouwkeet maar neemt een kopie mee zodat de originelen intact blijven. Zodra er voldoende eiwit af is, worden de kopietjes weggegooid. Op het moment dat er weer nieuw eiwit aangemaakt moet worden, worden er nieuwe kopieën gemaakt. Aan de aantallen kopieën van genen die aanwezig zijn in de cel, kunnen we aflezen welke eiwitten er op dat moment geproduceerd worden. Al in de eerste experimenten die mRNA kopietjes (expressie noemen we dat) maten vond men dat veel eiwitten die betrokken waren bij hetzelfde proces, vergelijkbare expressie niveaus hadden tijdens een celdeling.

We kunnen cellen ook manipuleren, bijvoorbeeld door het dieet aan te passen. Na het weglaten van een bepaalde bouwstof, kunnen we meten van welke genen hun expressie veranderd is. Van deze genen verwachten we dat ze betrokken zijn bij de aanmaak van deze bouwstof. Dit is al één manier om achter de functie van onbekende genen te komen. Een andere methode om de cel te manipuleren is door een bepaald gen uit het genoom weg te halen door gentechnologie. Nu is het niet meer gelijk bekend waar de genen die hun expres-

sie veranderen bij betrokken zijn maar het is wel mogelijk om groepen van genen te identificeren die steeds op dezelfde manier reageren op het weghalen van verschillende genen. Als het nu van een gen al bekend was wat het deed, kunnen we zeggen dat de andere genen van de groep waarschijnlijk bij hetzelfde proces betrokken zijn.

Nu hebben biologen heel veel genomics data verzameld over de expressie van alle genen op mRNA- maar ook op eiwit-niveau en alle deze data willen we natuurlijk gebruiken om iets te kunnen zeggen over de functies van al die onbekende genen. In dit proefschrift heb ik methoden ontwikkeld die data over genen uit hetzelfde organisme, maar ook uit verschillende organismen samenbrengen. Bijvoorbeeld als twee genen steeds tegelijkertijd tot expressie komen in gist maar ook in worm is het heel waarschijnlijk dat de gecodeerde eiwitten direct aan elkaar binden of dat ze samen een biologische rol vervullen. Ook heb ik verschillende soorten data bij elkaar gebracht om achter de functies van onbekende genen te komen. Het resultaat van het combineren van verschillende data, is dat de voorspelde functionele relaties veel betrouwbaarder zijn dan als we maar een enkele dataset gebruiken. Ook hebben we laten zien dat door verschillende data handig te combineren, we kunnen voorspellen of de gecodeerde eiwitten echt aan elkaar binden of dat ze samen in een proces hun functie vervullen. Het is belangrijk om de type relatie tussen eiwitten te specificeren, zodat er specifieke experimenten opgezet kunnen worden om te verifiëren dat de relatie ook daadwerkelijk bestaat.

De groepen genen met gezamenlijke expressie-profielen kunnen we ook nog ergens anders voor gebruiken. Genen worden bestuurd door regulerende eiwitten. Deze eiwitten binden aan het genoom, in de buurt van waar het gen gecodeerd is, en kan op deze manier beïnvloeden of het gen tot expressie komt. Ieder regulerend eiwit bindt aan een eigen specifieke volgorde van nucleotiden. In groepen genen met hetzelfde expressie-profiel in de malaria parasiet zijn we gaan zoeken naar combinaties van nucleotiden die in de buurt van alle genen van de groepen voorkwamen om er zo achter te komen, welke plekken in het genoom belangrijk zijn voor de regulatie van deze genen. We hebben daarbij niet alleen naar de omgeving op het genoom van de genen zelf hebben gekeken, maar tegelijkertijd naar de omgeving van dezelfde genen in het genoom van een andere malaria parasiet die knaagdieren infecteert. Van de stukjes genoom die belangrijk zijn voor de regulatie, verwachten we namelijk dat ze ook aanwezig zijn in het verwante organisme. Van de stukken genoom die minder belangrijk zijn omdat de specifieke volgorde van de nucleotiden geen functie heeft, verwachten we dat ze niet bewaard zijn gebleven maar veranderd door mutaties. Door naar twee verwante organismen tegelijk te kijken, hebben we specifieke combinaties van nucleotiden gevonden die belangrijk zijn voor gen-regulatie in de malaria parasiet. Welke regulerende eiwitten eraan binden en hoe de regulatie in zijn werk gaat, moet nu onderzocht worden.

We kunnen concluderen dat genomics data zeer nuttig is om inzicht te krijgen in de functies van genen en in de relaties tussen genen. Het combineren van verschillende typen data en data uit verschillende organismen is essentieel om hoogwaardige voorspellingen te kunnen doen over nieuwe relaties. Genomics data kan verder gebruikt worden om meer inzicht te krijgen in de regulatie van genen. Verder zal het ontwikkelen van methoden om genomics data te analyseren nodig blijven, omdat er steeds meer nieuwe en nieuwe soorten data geproduceerd worden.

Dankwoord

In het voorjaar van 2001 deed ik een onderzoeksproject bij David Ussery aan de Deense Technische Universiteit vlakbij Kopenhagen. Het instituut organiseerde in dat jaar ISMB, een van de grootste bioinformatica conferenties ter wereld. Tijdens deze conferentie presenteerde ik een poster over gen-annotatie in *E. coli*, waar onder andere Frank van Enckevort langskwam om het over mijn project te hebben; hij vertelde dat hij in Nijmegen bij het Centrum voor Moleculaire en Biomoleculaire Informatica werkte. Het leek mij de moeite waard om te onderzoeken of ik daar kon promoveren. Dus nam ik contact op met Gert Vriend van het CMBI en mocht een presentatie komen houden over mijn stage-project. Na gesprekken met zowel Martijn Huynen als Roland Siezen, besloot ik om bij de eerste onderzoek te gaan doen. Vanaf maart 2002 zaten we met zijn drieën op een kamer; Toni, Martijn en ik. Dat beviel uitstekend, want de samenwerking en de begeleiding was op deze manier zeer intensief. Martijn, als promotor heb ik heel veel aan jou gehad. Een paar keer per week vragen naar de resultaten is voor mij een goede stimulans geweest. Je vroeg vaak aan het eind van de dag nog de laatste versie van een manuscript om er 's avonds of in het weekend nog even naar te kijken. Verder was jij kritisch op momenten dat ik de resultaten eigenlijk al zeer veelbelovend vond. Ik wil jou dan ook als eerste heel erg bedanken. Toni, bedankt voor alle bomen. Toen de groep begon te groeien, verhuisde ik naar een kamer met Bas en Berend en dit beviel ook weer uitermate goed. We voerden 'vrijdagmiddag-gat-in-de-markt' in, maar helaas zonder enig resultaat. Wel waren we erg goed in het brainstormen over het onderzoek zelf, en de co-auteurschappen vooral samen met Berend zijn het bewijs dat dit nog effectief was ook. Daarnaast kletste ik natuurlijk ook gewoon veel te veel, dus heel veel dank voor het verdragen hiervan. Fiona, dat wij naast elkaar zaten verminderde de tijd die er gekletst werd uiteraard niet, maar het werd er wel gezelliger op. In de nieuwbouw samen jou en met Philip is er vooral veel zinnige discussie over wetenschap en politiek, dus jij en Philip bedankt. Daarnaast wil natuurlijk nog iedereen die bij het CMBI werkt en in het bijzonder de comics groep bedanken voor een hele fijne tijd. Koffiepauzes op vaste tijden waren altijd een gelegenheid om het nog even over iets heel anders te hebben dan wetenschap.

En dan zijn er alle vrienden. De eerste twee jaar van mijn promotie woonde ik nog in Utrecht en speelde bij het Utrechtsch Studenten Concert. De muziek is voor mij enorm belangrijk geweest als ontspanning en uitlaatklep. Iedereen van het orkest maar Karen en Reinie, hoorngenoten, wil ik in het bijzonder bedanken voor alles wat we meegemaakt hebben; de mooie (en soms ook minder mooie) muziek die we samen gemaakt hebben, de jaarlijkse tournees en de honorairen-installaties waar ik hier verder geen details over zal verstrekken. Ook de leden en in het bijzonder de hoornisten van Philharmonie Gelre en ook Mireille, Annet, Sander en Jaap van het blaaskwintet, bedankt voor de muziek.

Lieve vriendinnen van het Fioretti College, Wendy, Françoise, Debby, Marie-Louise, Linda, Annemiek en Suzanne, jullie wil ik bedanken voor het blijven onderhouden van het contact, ondanks dat ik daar niet het meest actief in ben geweest, en vooral voor de vriendinnenweekenden die we de laatste jaren hebben ingevoerd. Ik zie uit naar de dag dat de tijdscapsule open mag. Krokodil!

Daarnaast wil ik natuurlijk mijn ouders bedanken voor het altijd voor mij klaar staan, ook al snapten ze geen bal van wat ik nou eigenlijk aan het doen was en natuurlijk mijn lieve zussen, Guda en Jeannette, hoewel ver weg, altijd klaar op de MSN en op hyves met advies en goede raad. Michiel, als ik het niet meer zag zitten, wist jij me altijd weer op te beuren. Daarvoor en voor al die andere dingen wil ik je bedanken. En als laatste nog Raphael, mijn kleine engel, bedankt voor alle knuffels en kusjes.

Bedankt!

Vera

Publications

van Noort, V., Snel B. and Huynen M.A. Exploration of the omics evidence landscape to distinguish metabolic from physical interactions. *Submitted*.

van Noort, V. and Huynen, M.A. (2006) Combinatorial gene regulation in Plasmodium falciparum. *Trends Genet*, 22, 73-78.

Huynen, M.A., Snel, B. and **van Noort, V.** (2004) Comparative genomics for reliable protein-function prediction from genomic data. *Trends Genet*, 20, 340-344.

Snel, B., **van Noort, V.** and Huynen, M.A. (2004) Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. *Nucleic Acids Res*, 32, 4725-4731.

van Noort, V., Snel, B. and Huynen, M.A. (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep*, 5, 280-284.

van Noort, V., Snel, B. and Huynen, M.A. (2003) Predicting gene function by conserved co-expression. *Trends Genet*, 19, 238-242.

van Noort, V., Worning, P., Ussery, D.W., Rosche, W.A. and Sinden, R.R. (2003) Strand misalignments lead to quasipalindrome correction. *Trends Genet*, 19, 365-369.

Kesmir, C., **van Noort, V.**, de Boer, R.J. and Hogeweg, P. (2003) Bioinformatic analysis of functional differences between the immunoproteasome and the constitutive proteasome. *Immunogenetics*, 55, 437-449.

Curriculum vitae

Vera van Noort werd geboren op 26 augustus 1979 te Groningen. Vanaf augustus 1991 was zij leerling aan het Fioretti College te Lisse, waar zij in juni 1997 het Gymnasium diploma behaalde. In september 1997 begon zij met de studie biologie aan de Universiteit Utrecht, waar zij haar Propedeuse haalde met het predicaat Cum Laude. Tijdens haar studie deed Vera ervaring op in organisatie, beleid en bestuur door de uitvoering van verschillende bestuursfuncties bij het Utrechtsch Studenten Concert en het Koordinerend Orgaan van Studenten Muziekgezelschappen Utrecht. Zij deed afstudeeronderzoek bij Prof. dr. P. Hogeweg van de vakgroep Theoretische Biologie / Bioinformatica, Universiteit Utrecht. Dit onderzoek betrof de evolutie van het proteasoom. Een tweede afstudeeronderzoek werd gedaan bij dr. D.W. Ussery van het Center for Biological Sequence Analysis (CBS), Danmarks Tekniske Universitet (DTU), waarbij de genannotatie van *E. coli* werd bestudeerd. Verder schreef zij over het voorspellen van functionele associaties een scriptie bij Prof. dr. M.A. Huynen van het Centrum voor Moleculaire en Biomoleculaire Informatica (CMBI), UMC St Radboud. In februari 2002 haalde zij haar doctoraal examen biologie met als specialisatie Theoretische Biologie en Bioinformatica, waarna zij in maart direct begon met haar promotie onderzoek bij Prof. dr. M.A. Huynen. Onder zijn begeleiding verrichtte zij onderzoek waarvan de belangrijkste resultaten staan beschreven in dit proefschrift.